UNIVERSIDAD AUTÓNOMA DE MADRID
FACULTAD DE PSICOLOGÍA

Detección del Funcionamiento Diferencial del Ítem en Test Adaptativos Informatizados

Tesis Doctoral

Fabiola González Betanzos

Madrid, 2011

UNIVERSIDAD AUTÓNOMA DE MADRID

Facultad de Psicología

Departamento de Psicología Social y Metodología

Detección del Funcionamiento Diferencial del Ítem en Test Adaptativos Informatizados

Tesis Doctoral

Autora:

Fabiola González Betanzos

Directores

Francisco J. Abad García

Juan Ramón Barrada González

AGRADECIMIENTOS

Es muy difícil señalar a todas las personas que han contribuido para que yo pudiera concluir este trabajo, quisiera agradecer a todas aquellas que hicieron posible que una mexicana cogiera un buen día el avión y viniera a España a "obtener" un doctorado, con la promesa de regresar con el título apostillado bajo el brazo, nadie me advirtió que pasaría si me llevaba algo más…

Fue una fortuna para mi ser aceptada en el doctorado en Metodología que comparten tres de las universidades con más prestigio en España. Quiero agradecer a Jesús Alvarado, quien fue la primera persona que me abrió las puertas, primero de su despacho en la Universidad Complutense y después del Instituto de estudios biofuncionales, ahí compartí el aprendizaje y las comidas con Iwin y con Pablo, con ellos he aprendido que la amistad no tiene sus fronteras en el tiempo y en el espacio, sino en la nobleza y la lealtad. También recuerdo con afecto al profesor José Manuel Reales, su recibimiento y la empatía del personal de la UNED me dio la tranquilidad cuando más la necesitaba. Durante el segundo año recibí una invitación para asistir a un seminario en la Universidad Autónoma de Madrid, donde conocí a Francisco J. Abad, que me aceptara como estudiante ha sido de las mejores cosas que me han podido pasar, su paciencia y su trabajo incansable han permitido que yo haya podido concluir la tesis en las mejores condiciones. Quiero agradecer a Vicente Ponsoda a Julio Olea y a Paco por su generosidad en todos los sentidos, por mostrarme lo valioso de trabajar con un grupo de investigación de calidad. A Juanrra por venir cuando el camino se hacía más difícil y tender los puentes, por sus palabras.

El trabajo académico se lleva mucho mejor cuando hay personas a tu lado como Luis, Graciela, Sonia, Lara, Guaner. A Luis por su compañía, a Graciela, Sonia y Gloria que son mis mejores amigas en Madrid, gracias a Lara que me ayudó a iniciar la parte de simulación de los datos de la tesis, con Guaner he aprendido mucho y a Blanca que estuvo conmigo cuando más la necesitaba.

A México me voy con una tesis y con José Luis, un hombre con el valor suficiente para enfrentarse a un mundo desconocido, a él por la ilusión de vivir a su lado. A Isabel, su madre por darme una familia (Ma. Jesús, los tíos: Rosa, Natalia y Pedro).

Quiero agradecer a Mario Orozco el director de la Facultad de Psicología de la UMSNH por creer en mi, por su amistad, por construir una facultad plural.

A mis compañeras de la Facultad de psicología de la Universidad Michoacana de San Nicolás de Hidalgo, por escribirme y pensar en mi.

A Jennifer y Zaira por ser mucho más que mis mejores amigas, por nuestros sueños y proyectos compartidos, por estar siempre cerca de pesar de la distancia.

A Ulises, por ayudarme a ser quien soy

A Rocío y Rafael mis compañeros en las pequeñas y en las grandes batallas, por el ombligo que nos une al mundo.

A Jessica, Gabriel, Tania a María y el bebé en camino, por la esperanza

A mis padres… por estar siempre, porque mis logros les recuerden aquella primera alegría, Por fortalecerme con sus palabras, por todo ello estan eximidos de seguir leyendo.

Índice

Prefacio ..... **IX**
1. Introducción al Funcionamiento Diferencial del Ítem **1**
1.1. Introducción 1
1.2. El funcionamiento diferencial del ítem: Definición y tipos 2
1.3. Métodos de detección del DIF 5
1.4. Métodos UCI de detección del DIF basados en la TRI 8
1.4.1. El test IRT LR 8
1.4.2. Medidas de tamaño del efecto desde la TRI 9
1.4.3. Funcionamiento diferencial de los ítems y del test (DFIT) 11
1.5. Métodos de detección del DIF no basados en la TRI 11
1.5.1. Procedimiento de estandarización 12
1.5.2. SIBTEST (Simultaneous Item Bias Test) 13
1.5.3. Mantel-Haenszel 14
1.5.4. Regresión logística 16
1.6. Resultados de comparación de los métodos 17
2. Tests Adaptativos Informatizados **21**
2.1. Tests adaptativos informatizados 21
2.2. Componentes fundamentales de un TAI 23
2.3. Estimación del nivel de habilidad 24
2.4. Evaluación del criterio de parada 26
2.5. Criterios de selección de ítems 28
2.5.1. Estrategias de selección de ítems 29
2.5.2. Especificaciones de contenido en los test 31
2.5.3. Control de la exposición y seguridad del test 33
2.6. La evaluación de la fiabilidad y la validez en los TAIs 40
3. El estudio del DIF en Test Adaptativos Informatizados **43**
3.1. Introducción 43
3.2. Métodos adaptados para la evaluación de DIF en TAIs 46
3.2.1. Métodos basados en la TRI: IRT LR test 46

3.2.2. Métodos OCI aplicados a la detección del DIF en TAIs 47
3.2.3. Comparación de métodos 51
3.3. Conclusiones 56
4. Calibración online y análisis del DIF en ítems pretest 59
4.1. Introducción 59
4.2. Aplicación del algoritmo EM en la estimación por MML 61
4.3. La aplicación del algoritmo EM a un modelo multigrupo 62
4.4. Restricciones para la identificación en un modelo multigrupo 65
4.5. Calibración con Parámetros Fijos (CPF) 66
4.5.1. No actualización de la distribución previa y un ciclo EM (NWU-OEM) 69
4.5.2 No actualización de la distribución previa y múltiples ciclos EM (NWU-MEM) 70
4.5.3. Una sola actualización de la distribución previa y un ciclo EM (OW-OEM) 70
4.5.4. Una sola actualización de la distribución previa y Múltiples ciclos EM (OWU-MEM) 71
4.5.5. Múltiples actualizaciones de la distribución y múltiples ciclos EM (MWU-MEM) 72
4.6. Estudios previos sobre la calibración de ítems con parámetros fijos 73
4.7. Aplicación de los métodos de CPF al análisis del DIF 75
4.8. Anexo 77
4.8.1. Ignorabilidad de los valores perdidos en un TAI 77
5. Estudio 1: Detección online del DIF con métodos de CPF 81
5.1. Abstract 81
5.2. Calibración online en TAIs 82
5.3. Detección del DIF en TAIs 85
5.4. Detección online del DIF en TAIS con métodos de CPF 86
5.5. Método 88
5.5.1. Condiciones de aplicación del TAI 88
5.5.2. Parámetros de los ítems operativos 88
5.5.3. Parámetros de los ítems pretest 89
5.5.4. Factores manipulados 90
5.5.5. Métodos de calibración 91
5.5.6. Detección del DIF 92

5.5.7. Criterios de valoración ... 92
5.5.8. Análisis ... 93
5.6. Resultados ... 94
5.6.1. Detección del DIF ... 94
5.6.2. Efectos del método en la recuperación de la escala ... 97
5.6.3. Recuperación de los parámetros de los ítems pretest ... 99
5.6.4. Recuperación del Tamaño del DIF ... 100
5.7. Discusión y conclusiones ... 113
5.8. Referencias ... 116
6. Estudio 2: Evaluación del DIF en ítems aplicados mediante un TAI: IRT LRT vs CATSIB ... 119
6.1. Abstract ... 119
6.2. IRT LRT ... 122
6.3. CATSIB ... 123
6.4. Tamaño del DIF en TAIs ... 125
6.5. Método ... 126
6.5.1. Condiciones de aplicación del TAI ... 126
6.5.2. Banco de ítems ... 127
6.5.3. Ítems analizados ... 127
6.5.4. Factores manipulados ... 128
6.5.5. Criterios de valoración ... 130
6.5.6. Análisis ... 131
6.6. Resultados ... 132
6.6.1. Tasas de error Tipo I en la condición sin DIF (Banco no contaminado) 132
6.6.2. Tasas de error tipo I en las condiciones con DIF (Banco contaminado) 135
6.7. Discusión y conclusiones ... 141
6.8. Referencias ... 144
7. Conclusiones y Discusión General ... 149
7.1. Consideraciones Generales ... 149
7.2. Limitaciones de los estudios y futuras investigaciones ... 152
7.3. Conclusiones ... 157
Referencias ... 159

Índice de Tablas

Tabla 4.1. Marco para la clasificación de los métodos de CPF (Kim, 2006). 67
Tabla 5.1. Parámetros para generar los ítems en los que se estudia el DIF 89
Tabla 5.2. Tasas de error Tipo I para todas las combinaciones de las condiciones en el estudio 1 94
Tabla 5.3. Tasas de Potencia para el DIF unidireccional y no-unidireccional 95
Tabla 5.4. Medias y desviaciones típicas promedio de la distribución de la habilidad para el grupo de referencia y el grupo focal 98
Tabla 5.5. Media de Sesgo y RMSE en la estimación del tamaño del efecto (ausencia de DIF) 101
Tabla 5.6. Media de Sesgo y RMSE en la estimación del tamaño del efecto (DIF unidireccional) 102
Tabla 5.7. Media de Sesgo y RMSE en la estimación del tamaño del efecto (DIF no unidireccional) 105
Tabla 6.1. Parámetros de los ítems estudiados en el Banco contaminado 128
Tabla 6.2. Tasas de error Tipo I para los conjuntos de datos libres de DIF. Ítems del 271 al 285 133
Tabla 6.3. Tasas de error Tipo I para los conjuntos de datos libres de DIF (banco no contaminado). Ítems del 286 al 300 134
Tabla 6.4. Tasas de error Tipo I para los conjuntos de datos con DIF. Ítems del 271 al 285. 136
Tabla 6.5. Tasas de Potencia para los ítems con DIF. 137
Tabla 6.6. Sesgo y RMSE promedio para el tamaño del DIF por impacto y tamaño de las muestras en cada método. 139

Índice de Figuras

Figura 1.1. Representación gráfica del DIF unidireccional (izquierda) y DIF no unidireccional (derecha). 5
Figura 2.1. Diagrama de flujo de un TAI (tomado de Barrada, en prensa). 24
Figura 5.1. Potencia para cada método, en función del impacto y la longitud del TAI . 96
Figura 5.2. Sesgo en la estimación de β cuando el DIF es unidireccional, como función de método, impacto y longitud del TAI..... 103
Figura 5.3. Sesgo y RMSE en la estimación del parámetro *a* del grupo de referencia como función del método, impacto y longitud del TAI y tamaño de las muestras 107
Figura 5.4. Sesgo y RMSE en la estimación del parámetro *b* del grupo de referencia como función del método, impacto y longitud del TAI y tamaño de las muestras 108
Figura 5.5. Sesgo y RMSE en la estimación del parámetro *c* del grupo de referencia como función del método, impacto y longitud del TAI y tamaño de las muestras 109
Figura 5.6. Sesgo y RMSE en la estimación del parámetro *a* del grupo de focal como función del método, impacto y longitud del TAI y tamaño de las muestras 110
Figura 5.7. Sesgo y RMSE en la estimación del parámetro *b* del grupo de focal como función del método, impacto y longitud del TAI y tamaño de las muestras 111
Figura 5.8. Sesgo y RMSE en la estimación del parámetro *c* del grupo de focal como función del método, impacto y longitud del TAI y tamaño de las muestras 112
Figura 6.1. Tasa de error Tipo I local según método de detección de DIF e impacto.. 135
Figura 6.2. Sesgo en la estimación de β en la condición con DIF, como función del impacto 140
Figura 6.3. RMSE en la estimación de β en la condición con DIF, como función del impacto 141

Prefacio

En cualquier proceso de evaluación psicológica es importante estudiar que las puntuaciones obtenidas por las personas no tienen *sesgo*. El sesgo se entiende en este contexto como un error sistemático relacionado con el efecto de alguna variable asociada a la pertenencia del sujeto a un determinado grupo demográfico y que es ajena al atributo psicológico que se mide. La estrategia habitual para evaluar la presencia de sesgo es estudiar el funcionamiento diferencial de los ítems (DIF). Así, los estándares para la construcción y validación de tests (AERA, APA, NCME, 1999) señalan que las organizaciones están *obligadas* a realizar pruebas de DIF a través de las posibles subpoblaciones de evaluados, con el fin de detectar y remover o revisar los ítems que presentan DIF. Esta exigencia no debe ser infravalorada, especialmente en contextos en los que la puntuación en el test tenga consecuencias importantes (laborales, educativas, clínicas, etc.) y se desee garantizar la igualdad de oportunidades. Por esta razón, en las tres últimas décadas, se han desarrollado numerosos métodos para el análisis del DIF en test convencionales.

Paralelamente, en los años 90 del siglo XX, la sinergia entre el desarrollo informático y el psicométrico ha revolucionado los procedimientos de evaluación, hasta el punto que Embretson (2004) distingue entre tests de primera generación y los nuevos tests, de segunda generación, entre los que se incluyen los tests adaptativos informatizados (TAIs). Desafortunadamente, los TAIs permiten optimizar el proceso de evaluación pero suponen cambios, en la forma de aplicación de los ítems y de asignar puntuaciones, que dificultan la aplicación de los métodos tradicionales de análisis del DIF. En el presente trabajo de tesis doctoral se describen algunas de las aproximaciones al estudio del DIF en TAIs y se propone un nuevo procedimiento, inspirado en los recientes desarrollos obtenidos en el ámbito de la calibración online. En concreto, nuestro objetivo principal es obtener una adaptación válida del Test de Razón de

Verosimilitudes de la TRI (IRT LRT) que permita realizar el análisis del DIF tanto en ítems pretest, como en ítems operativos.

La tesis se estructura en siete capítulos. En el primer capítulo se establecen algunas definiciones básicas en relación a este ámbito de investigación y se describen aquellos métodos, paramétricos y no parámetricos, propuestos para detectar DIF en tests convencionales (Camilli y Shepard, 1994; Holland y Wainer, 1993; Osterlind y Everson, 2009). Se realiza también la distinción entre los métodos de DIF condicionados a la variable observada (OCI, por sus siglas en inglés) y los métodos de DFI condicionados a la variable latente (UCI). La descripción no pretende ser exhaustiva, ya que se centra en los métodos que posteriormente se han adaptado al análisis del DIF en TAIs (i.e. Mantel-Haenzsel, el método de estandarización, la regresión logística, SIBTEST e IRT LRT).

En el capítulo segundo se explica la estructura y el funcionamiento de los TAIs, que se fundamenta en los modelos de teoría de respuesta al ítem (TRI). Se describen los principales procedimientos de selección de ítems y de control de la exposición. Además, se establecen algunas de las potenciales amenazas a la validez que son propias de los TAIs (p. ej., cambios en los parámetros de los ítems debidas a la difusión de los contenidos del banco) y que implican la necesidad de controles psicométricos continuos durante las distintas fases de la vida operativa del programa.

En el capítulo tercero exponemos los problemas y las formas en las que algunos métodos se han adaptado para el estudio del DIF en TAIs. Los problemas asociados al estudio del DIF dependen de la fase operativa. En la fase de construcción del banco, las aplicaciones de los ítems no suelen ser adaptativas y se aplican los procedimientos de detección del DIF clásicos. Por el contrario, en las fases de actualización y mantenimiento algunos ítems se aplican adaptativamente. En la fase de actualización, el objetivo es detectar el DIF de los ítems diseñados para actualizar el banco (denominados ítems *pretest*). Los ítems pretest se aplican a una muestra amplia, en la que a la vez que se aplican, de forma adaptativa, los ítems del TAI (denominados ítems operativos). En la fase de mantenimiento, se estudia el DIF de los ítems operativos, para garantizar que sus propiedades psicométricas continúan siendo aceptables a pesar de su aplicación continuada. En ambas fases, surgen dos dificultades importantes: (a) la ausencia de una variable de igualación adecuada para los métodos OCI (la suma de respuestas correctas no es interpretable en un TAI); (b) la ingente cantidad de valores perdidos en la matriz de datos, que dificulta la aplicación de los métodos UCI.

Así, en relación a los procedimientos OCI, el problema suele reducirse a encontrar una variable de igualación adecuada. Por ejemplo, en la adaptación de SIBTEST, CATSIB, se utiliza el nivel de habilidad estimado en el TAI como variable de igualación (Nandakumar y Roussos, 2001, 2004). En relación al procedimiento UCI estudiado en esta tesis, IRT LRT, el problema es superar la presencia de valores perdidos que introducen problemas en la re-calibración de los parámetros de los ítems operativos. La única solución propuesta ha sido imputar las respuestas faltantes a partir del nivel de habilidad estimado en el TAI (Lei, Chen y Yu, 2006), pero esta propuesta resulta ineficiente en términos de coste computacional y tiene algunas complicaciones asociadas al software disponible, IRTLRDIF (Thissen, 2001).

El capítulo 4 es el núcleo teórico de la tesis. El problema de tener que re-calibrar los parámetros de los ítem operativos con pocos sujetos, ya ha sido investigado previamente en el contexto de la calibración online. La calibración online se refiere al proceso mediante el que los parámetros de los ítems pretest se estiman y se sitúan en la misma escala métrica de los ítems operativos por su calibración conjunta, en el que los ítems operativos del TAI se definen como ítems de anclaje. Wainer y Mislevy (1990) fueron los primeros en reconocer que en este contexto era prescindible volver a calibrar los ítems operativos, puesto que ya se conocían. A partir de esta idea surgen los métodos de calibración con parámetros fijos (CPF), cuya característica fundamental es que se fijan los parámetros de los ítems operativos a sus valores conocidos y se estiman únicamente los ítems pretest. Existen distintas estrategias, descritas en Kim (2006), que se recogen en el cuarto capítulo. Mediante estos procedimientos desaparece el problema de la falta de datos en los ítems operativos. En concreto, puesto que ya no recalibran los ítems operativos, se eliminan los problemas de convergencia asociados a su estimación y se disminuye el tiempo de calibración online. Además, los parámetros de los ítems pretest se obtienen en la escala métrica de los ítems operativos.

Al final del capítulo 4, se describe la adaptación de la prueba IRT LRT, haciendo uso de los métodos CPF. Se presentan algunas singularidades de dicha aplicación, entre ellas las restricciones necesarias para estimar estos modelos. Por último, se describen las hipótesis a las que da lugar la aplicación de los métodos de CPF al análisis del DIF.

En el estudio 1 se ponen a prueba las hipótesis propuestas en el capítulo anterior en un contexto de calibración online o detección del DIF en ítems pretest. Se aplicaron tres métodos de CPF, aquellos en los que se actualizan las distribuciones de la

habilidad, y se compararon con la propuesta de imputación de Lei (Lei, et al., 2006). En el estudio 2 se estudian los métodos en el contexto de la detección del DIF en ítems operativos. El mejor método CPF, según los resultados del estudio 1, se compara con el método CATSIB (Nandakumar y Roussos, 2001, 2004). Los resultados obtenidos en ambos estudios sugieren que la propuesta de aplicar IRT LRT, conjuntamente con un método CPF, puede resultar un procedimiento más adecuado que imputar las respuestas (estudio 1) o aplicar CATSIB (estudio 2). Finalmente, en la discusión general se exponen los principales resultados, las limitaciones de los estudios realizados y las futuras líneas de investigación.

Capítulo 1

Introducción al Funcionamiento Diferencial del Ítem

1.1. Introducción

Los resultados de la evaluación a través de los test pueden tener consecuencias adversas en la vida cotidiana de las personas. Basta pensar en un contexto de selección laboral en el que una persona deja de ser contratada en base a los resultados de una prueba. Estas circunstancias comprometen a los científicos a realizar estudios que garanticen *la equidad* en la medición y la validez de las inferencias que se hacen a partir de las puntuaciones en los test (Borsboom, Romeijn y Wicherts, 2008). Es difícil defender un test que resulte *sesgado* contra un grupo en función de su etnia, sexo, cultura u otra característica sociodemográfica. El término *sesgado*, en este contexto, implica que las puntuaciones tienen distinto significado para miembros de diferentes grupos. Así, en su desarrollo, los test deben pasar por procesos de análisis del *sesgo* para evitar la discriminación a las minorías étnicas (afroamericanos, hispanos, asiáticos) o entre sexos, por poner un ejemplo.

Una aproximación básica al estudio del sesgo es el estudio del funcionamiento diferencial de los ítems o del test. El funcionamiento diferencial de un ítem (*differential item functioning*; DIF) o de un test (*differential test functioning*; DTF) se produce cuando el ítem (o el test) tiende a proporcionar puntuaciones distintas para personas que pertenecen a diferentes grupos y que, sin embargo, tienen el mismo nivel de rasgo. Estos términos (DIF, DTF) son semánticamente más neutrales que el término *sesgo*, cuyo uso suele restringirse a la explicación teórica de cuáles pueden ser las variables "*ajenas*" involucradas en la aparición del DIF (Shealy y Stout, 1993a; Fidalgo, 1996).

El término *impacto* se reserva para las diferencias reales entre los grupos (Camilli y Shepard, 1994).

En los siguientes apartados se proporciona una definición más precisa del DIF y se definen los tipos de DIF. A continuación se describen algunos de los métodos que se han propuesto para el estudio del DIF, haciendo especial énfasis en el test de razón de verosimilitudes de la teoría de la respuesta al ítem. Finalmente, se resumen algunos de los resultados principales obtenidos cuando se comparan estos métodos.

1.2. El funcionamiento diferencial del ítem: Definición y tipos

En términos generales podemos afirmar que el DIF ocurre cuando existen diferencias en las propiedades métricas de un ítem entre diversos grupos. En el estudio del DIF usualmente la comparación se realiza entre dos grupos; al grupo de interés se le llama *grupo focal*, y suele coincidir con el grupo minoritario, mientras que al grupo mayoritario se le llama *grupo de referencia.* Millsap y Everson (1993) proporcionan un marco general para definir formalmente el *funcionamiento diferencial* de una medida *Y,* distinguiendo tres elementos básicos:

a) Una variable latente, θ, que indica el nivel de rasgo y es el objeto de la medición.
b) Una variable observable, *Y*, que es un indicador de θ (p.e., puede ser la puntuación en un ítem o en el test).
c) La variable de agrupación, *g*, que define a los grupos, generalmente diferenciados en base a algún tipo de información socio-demográfica (p.e., género, grupo étnico, edad, etc.).

A partir de estos elementos se considera que *no existe funcionamiento diferencial* si el valor esperado de la puntuación, *Y,* condicionado a la variable latente, θ, no depende del grupo, *g* (Millsap y Everson, 1993):

$$E(Y \mid \theta, g) = E(Y \mid \theta) \tag{1.1}$$

La presencia de funcionamiento diferencial de *Y* implica que, para algún nivel de θ:

$$E(Y \mid \theta, g) \neq E(Y \mid \theta) \tag{1.2}$$

Para entender las razones por las que puede encontrarse funcionamiento diferencial (FD) puede acudirse al marco teórico de la *teoría multidimensional del funcionamiento diferencial del ítem* (Shealy y Stout, 1993b). Consideremos que la medida *Y* es la puntuación en el ítem, u_j. En este marco, se establece que u_j es un indicador de la dimensión principal, θ, pero también de una o más dimensiones secundarias (p.e., ε). La puntuación esperada en u_j condicionada a θ se escribe como:

$$E(\mathrm{u_j} \mid \theta, g) = \int E(\mathrm{u_j} \mid \theta, \varepsilon, g) g(\varepsilon \mid \theta, g) d\varepsilon \tag{1.3}$$

Por tanto, en presencia de multidimensionalidad, puede darse ausencia de DIF si se cumple simultáneamente que:

1) Los parámetros son invariantes a través de los grupos:

$$E(\mathrm{u_j} \mid \theta, \varepsilon, g = R) = E(\mathrm{u_j} \mid \theta, \varepsilon, g = F) \tag{1.4}$$

2) La distribución condicionada de ε es la misma en los dos grupos

$$g(\varepsilon \mid \theta, g = R) = g(\varepsilon \mid \theta, g = F) \tag{1.5}$$

La presencia de ε no garantiza por sí sola la aparición del DIF ya que éste se manifestará sólo cuando las distribuciones condicionadas de los rasgos espurios (ε) sea diferente a través de los grupos (Oshima y Miller, 1992).

El modelo multidimensional proporciona un marco conceptual desde el que comprender la presencia de DIF. Este marco es interesante porque permite entender por qué la presencia de DIF puede variar de muestra a muestra. No obstante, resulta difícil de aplicar ya que la estimación de los parámetros del modelo multidimensional no es factible si el rasgo espurio afecta específicamente a un único ítem. Por ello, las técnicas usuales para la detección del DIF parten directamente de un modelo unidimensional, sin considerar explícitamente la presencia de rasgos espurios.

Por ejemplo, para ítems dicotómicos, se define la presencia de DIF cuando, al aplicar un modelo unidimensional, la probabilidad de acierto condicionada a θ difiere a través de los grupos:

$$P(u_j = 1 \mid \theta, g = R) \neq P(u_j = 1 \mid \theta, g = F) \tag{1.6}$$

En el marco de la TRI, la función de probabilidad de acierto condicionada se denomina curva característica del ítem (CCI) y se ajusta a una función predeterminada. Por ejemplo, en el modelo logístico de 2 parámetros (2PL):

$$P(u_j = 1 \mid \theta, g) = \frac{1}{1 + e^{-1.7 a_{jg}(\theta - b_{jg})}} \tag{1.7}$$

donde a_{jg} y b_{jg} son los parámetros de discriminación y dificultad del ítem j en el grupo g. Desde una perspectiva de TRI, evaluar el DIF es equivalente a contrastar si los parámetros (o las CCI) difieren entre grupos.

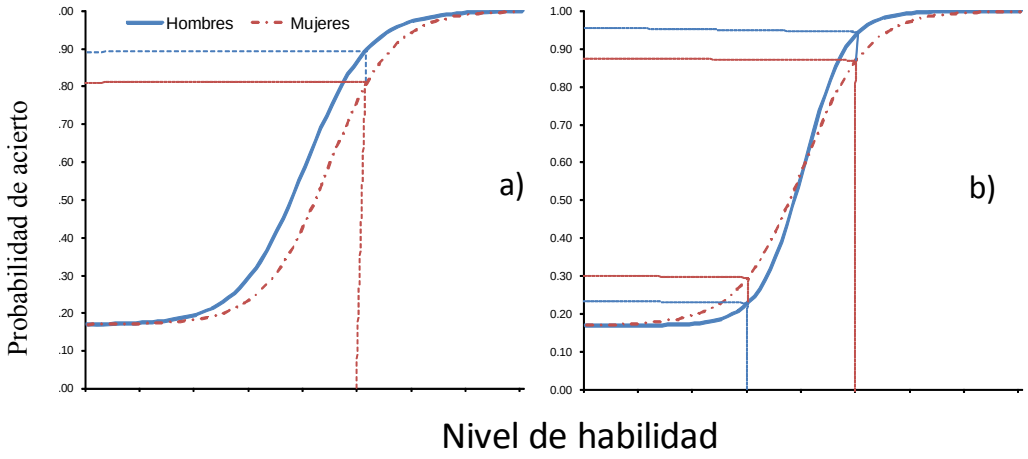
Suele distinguirse entre DIF *unidireccional* y *no unidireccional* (Hanson, 1998):

a) El DIF unidireccional implica que la diferencia entre los grupos ocurre en la misma dirección en todo el rango de θ. Eso ocurre cuando un grupo tiene una mayor probabilidad de acierto en todos los niveles. Un caso especial de DIF unidireccional es el DIF *paralelo*, que ocurre cuando en el modelo 2PL difieren los parámetros *b* pero no los parámetros *a*.

b) El DIF no unidireccional, por otro lado, se refiere al caso en el que las CCI se cruzan (el sentido de la diferencia entre las CCI de los grupos depende de θ); esto ocurre cuando el parámetro *a* difiere a través de los grupos (aunque el parámetro *b* también puede ser diferente); un caso especial de DIF no unidireccional es el DIF *simétrico*, que ocurre cuando solo difieren los parámetros *a*, pero no los parámetros *b.*

La distinción entre DIF unidireccional y no unidireccional se representa en la Figura 1.1 y es importante, ya que los distintos métodos de detección del DIF difieren en su potencia para detectar cada tipo de DIF. Por ejemplo, el DIF no unidireccional puede ser difícil de detectar en aquellos procedimientos en los que el indicador de DIF es un promedio de la discrepancia entre CCIs a través de θ.

Figura 1.1. Representación gráfica del DIF unidireccional (izquierda) y DIF no unidireccional (derecha).



1.3. Métodos de detección del DIF

Como ya se ha mencionado anteriormente, en cualquier método de detección del DIF se pretende contrastar si la puntuación esperada condicional difiere a través de los grupos. En cualquier método de detección del DIF se requiere el conocimiento previo de uno o más ítems *no sesgados*, cuyos parámetros se asumen invariantes a través de los grupos. A esos ítems se los denomina ítems de *anclaje*. Conocido el subconjunto de ítems en el que las CCI son iguales, es posible definir la variable condicional o variable de *igualación.* Las técnicas para la detección del DIF se pueden dividir según dos criterios:

1) La relación que se establece entre la puntuación en el ítem y la variable de igualación (Hidalgo, Gomez-Benito, 2010), que puede seguir un modelo paramétrico (p.ej., logístico) o un modelo no paramétrico. La aplicación de un modelo paramétrico supone ventajas pues permite extrapolaciones que serían imposibles siguiendo un modelo no paramétrico. Sin embargo, puede suponer un riesgo si las respuestas no se ajustan al modelo aplicado.
2) El modo en que se establece la variable condicional:

2.1. Los métodos que utilizan la puntuación total en el test (X) como estimación de θ, denominados también métodos de invarianza condicional observada (OCI). En estos modelos se contrasta si:

$$P(u_j \mid X, g = R) = P(u_j \mid X, g = F) \tag{1.8}$$

En los procedimientos OCI, se requiere el cálculo de la puntuación *X*, para posteriormente estudiar la relación entre *X* y u_j en ambos grupos. *X* es la variable de igualación.

2.2. Los métodos que evalúan el DIF condicionando directamente a la variable latente, llamados métodos de invarianza condicional no observada (UCI). En estos modelos se contrasta directamente si:

$$P(u_j \mid \theta, g = R) = P(u_j \mid \theta, g = F) \tag{1.9}$$

En los procedimientos UCI, θ es desconocida, pero, asumiendo un modelo matemático, se infiere la función de relación entre θ y u_j para cada grupo (i.e., un modelo de medida). En los métodos UCI paramétricos contrastar el DIF se reduce a evaluar la invarianza de los parámetros del modelo a través de los grupos.

La principal diferencia entre los métodos OCI y los métodos UCI es que en los primeros, ya sean paramétricos o no paramétricos, no se tiene en cuenta que la puntuación en el test *X* puede ser un mal indicador de θ (p.ej., si el test es corto). Por ello los métodos UCI paramétricos son teóricamente más apropiados en la detección del DIF (tasas de error tipo I más ajustadas y mayores tasas de potencia) aunque, a la vez, la presencia de un modelo de medida implica mayor exigencia (requieren estimar los parámetros de un modelo y comprobar el ajuste de los datos a éste).

Sea cual sea el método utilizado, pueden hacerse algunas consideraciones generales:

1) Debe contrastarse que los ítems de anclaje no tiene DIF. A la presencia de ítems con DIF en el test de anclaje se le denomina *contaminación*. Cuanto mayor sea la contaminación del test, mayor será la tasa de falsos positivos y menor será la potencia para detectar el DIF, independientemente del método empleado

(Clauser, Mazor y Hambleton, 1993; French y Maller, 2007; Fidalgo, Mellenbergh y Muñiz, 1998; 1999; 2000; Finch, 2005; Woods, 2008a). El problema de la contaminación se puede resolver incluyendo una etapa de depuración, en la que se eliminen los ítems con DIF en el primer análisis (Clauser et al., 1993; Holland y Thayer, 1988; Miller y Oshima, 1992), o varias etapas de depuración, estableciendo un procedimiento iterativo, hasta que se alcance la convergencia en la decisión en iteraciones sucesivas (Candell y Drasgow, 1988; Lautenschlager, Flaherty y Park, 1994; Lord, 1980). Algunas de las investigaciones se han centrado en cómo determinar los ítems que deben formar parte del test de anclaje (Woods, 2008a, González-Betanzos y Abad, en prensa).

2) En los métodos OCI debe incluirse el ítem estudiado en el cálculo de la puntuación en el test, independientemente de que tenga DIF o no. En presencia de impacto, los métodos OCI producen inflación de la tasa de error tipo I cuando no se incluye el ítem (Lewis, 1993; Zwick, 1990). Holland y Thayer (1988) demuestran formalmente que si las respuestas siguen el modelo de Rasch lo correcto es incluir el ítem. En ese caso, X es un estimador suficiente de θ. Esta recomendación también se ha hecho para los procedimientos de estandarización (Donoghue, Holland y Thayer, 1993) y la regresión logística (DeMars, 2009).
3) El resultado estadístico de ausencia/presencia del DIF debe acompañarse siempre de medidas de tamaño del efecto. Discrepancias pequeñas en las CCI pueden ser estadísticamente significativas si la muestra es grande; asimismo, grandes discrepancias pueden no ser estadísticamente significativas en muestras pequeñas.
4) En el análisis del DIF es importante contrastar si, al considerarse conjuntamente los ítems con DIF, el funcionamiento diferencial se amplifica o se cancela al nivel del test.
5) Se asume que los ítems con DIF son minoría o que hay un conjunto de ítems sin DIF previamente conocido. Es decir, se contrasta el funcionamiento diferencial de un ítem en relación a otros (que pueden ser la mayoría o el subconjunto elegido por el investigador). Por ejemplo, si todos los ítems tienen un funcionamiento diferencial tal que:

$$b_{jF} = b_{jR} + \gamma_j \tag{1.10}$$

La clasificación del DIF dependerá de qué ítems tomen parte del test de anclaje.

1.4. Métodos UCI de detección del DIF basados en la TRI

Se han propuesto diversas estrategias para contrastar si las dos CCI difieren significativamente, aunque el test de razón de verosimilitudes (IRT LR test) es el que se ha hecho más popular por la presencia de software fácil de usar (Thissen, 1991, 2001). Otras estrategias de TRI para contrastar la significación estadística son el χ^2 de Lord (Lord, 1980), las medidas de área con signo y sin signo entre dos CCI (Linn, Levine, Hastings y Wardrop, 1981; Raju, 1988, 1990; Rudner, Geston y Knight, 1980, Shepard, Camilli y Williams, 1984, 1985) o los índices de DIF compensatorio y no compensatorio (Raju, van der Linden y Fleer, 1995). Una ventaja de la TRI es que resulta fácil contrastar tanto el DIF unidireccional como el DIF no unidireccional.

1.4.1. El test IRT LR

El test IRT LR fue propuesto por Thissen, Steinberg y Gerrard (1986). Es un contraste general que implica una comparación de las CCI de los ítems para evaluar el DIF a través de los grupos y está basado en la diferencia en los estadísticos G^2 de dos modelos anidados: (1) un modelo restringido *m*', donde todos los parámetros se restringen a ser iguales a través de los grupos; y (2) un modelo aumentado *m*, que permite que los parámetros del ítem bajo estudio varíen a través de los grupos. Los estadísticos G_m^2 y $G_{m'}^2$ se definen como menos dos veces la log-verosimilitud de las respuestas. Para el modelo *m*:

$$G_m^2 = -2\sum_g \sum_i \ln \int \prod_j P(u_{ij} \mid \theta, \hat{\boldsymbol{\Delta}}_{\mathbf{m}}, g)\phi_g(\theta; \mu_g, \sigma_g)d\theta \tag{1.11}$$

donde u_{ij} es la respuesta del evaluado *i* al ítem *j*, $\hat{\boldsymbol{\Delta}}_{\mathbf{m}}$ es el vector con los parámetros de los ítems en el modelo *m* y $\phi(\theta; \mu_g, \sigma_g)$ expresa la distribución normal de θ con media μ_g y desviación típica σ_g para el grupo *g*. El estadístico $G_{m'}^2$ del modelo *m*' se define análogamente. En ambos modelos los parámetros de los ítems de anclaje se restringen a

ser iguales a través de los grupos (para enlazar la métrica del rasgo latente en los dos grupos), la media μ_g y la desviación típica σ_g para el grupo de referencia se fijan a 0 y 1 (para identificar la escala) mientras que la media μ_g y la desviación típica σ_g para el grupo focal se estiman libremente. Bajo la hipótesis nula (i.e., los datos se ajustan al modelo aumentado), la diferencia $G_m^2 - G_{m'}^2$ sigue una distribución χ^2 con l grados de libertad, donde l es el número de parámetros que están fijados en el modelo restringido y son libres en el modelo aumentado. Valores extremos para el estadístico son indicadores de que las CCI difieren significativamente a través de los grupos (Thissen, et al., 1986). Si el test estadístico general es estadísticamente significativo, posteriores análisis pueden llevarse a cabo para evaluar qué parámetros (*a, b* y/o *c*) difieren a través de los grupos).

1.4.2. Medidas de tamaño del efecto desde la TRI

Además de la significación estadística es importante obtener un indicador del tamaño del DIF que indique la relevancia del DIF. Steinberg y Thissen (2006) señalan que una posibilidad en el ámbito de la TRI es tomar las diferencias en el parámetro *b* entre los dos grupos, cuando el DIF es unidireccional este indicador está relacionado con las medidas de área.

Las medidas de área para el DIF expresan la diferencia entre las CCI del grupo de referencia y grupo focal como una función del área encerrada entre dichas funciones. Defínase la diferencia entre las CCI como:

$$D_{CCI}(\theta) = P(u_j = 1 \mid \theta, g = R) - P(u_j = 1 \mid \theta, g = F) \tag{1.12}$$

En las medidas de área se calcula (o estima) la integral de dicha función a través de θ. Por ejemplo, el área con signo (*signed area*; SA) entre dos ICCs (Raju, 1988, 1990; Rudner et al., 1980) puede obtenerse como:

$$SA = \int D_{CCI}(\theta) d\theta \tag{1.13}$$

Esta es una medida sensible al DIF unidireccional. El área sin signo (*unsigned area*; USA) puede obtenerse como:

$$USA = \int \left| D_{CCI}(\theta) \right| d\theta \tag{1.14}$$

que también es sensible a la presencia de DIF no unidireccional. Asumiendo que el parámetro *c* es idéntico a través de los grupos, Raju (1988) proporciona las ecuaciones para estimar SA y USA en los modelos de uno (1PL), dos (2PL) o tres parámetros (3PL). Por ejemplo, en el modelo de tres parámetros:

$$SA = (1-c)(b_F - b_R) \tag{1.15}$$

Esta formulación es independiente de los valores de *a*. Nótese que SA es proporcional a la diferencia entre parámetros *b*, lo que apoyaría la propuesta de Steinberg y Thissen (2006). El tamaño del efecto suele considerarse pequeño, moderado o grande según las áreas, en valor absoluto, se aproximen a 0.4, 0.6 o 0.8, respectivamente (Narayanan y Swaminathan, 1996; Rogers y Swaminathan, 1993). No obstante las medidas de área tiene una limitación importante: no consideran la distribución del rasgo en el grupo focal. En ese sentido, pueden ser mejores otras medidas de tamaño del efecto en las que $D_{CCI}(\theta)$ se pondere por la distribución de $f_j(\theta)$, que es la función de densidad del grupo total que responde al ítem *j* (Nandakumar y Roussos, 2001). Una medida que puede obtenerse es la propuesta en el procedimiento SIBTEST, β:

$$\beta_{j(SIBTEST)} = \int \left(P(u_j = 1 \mid \theta, g = R) - P(u_j = 1 \mid \theta, g = F)\right) f_j(\theta) d\theta \tag{1.16}$$

β: es sensible únicamente al DIF *unidireccional.* En el esquema de clasificación sugerido por Dorans (1989) se considera un DIF moderado si $0.05 \le \left|\beta_j\right| \le 0.10$ y un DIF grande si $\left|\beta_j\right| \ge 0.10$ (valores positivos indican que el ítem favorece al grupo de referencia). Roussos y Stout (1996) propone utilizar como puntos de corte 0.059 y 0.088, respectivamente.

Finalmente, otra propuesta es utilizar un índice análogo al propuesto por Raju, van der Linden y Fleer (1992, 1995), el índice de funcionamiento diferencial del ítem no compensatorio (*non-compensatory differential ítem functioning*; NCDIF):

$$NCDIF_j = \int \left(P(u_j = 1 \mid \theta, g = R) - P(u_j = 1 \mid \theta, g = F)\right)^2 f_j(\theta) d\theta \tag{1.17}$$

NCDIF es un promedio de la diferencia cuadrática media a través del nivel de rasgo y es sensible al DIF no unidireccional.

1.4.3. Funcionamiento diferencial de los ítems y del test (DFIT)

Raju, van der Linden y Fleer (1992, 1995) introdujeron un nuevo marco para evaluar el funcionamiento diferencial de los ítems y del test (DFIT). Se proponen medidas de discrepancia a nivel de ítem:

$$NCDIF_j = \mathrm{E}_F\left[D_{CCI_j}(\theta)\right]^2 \tag{1.18}$$

y a nivel de test:

$$DTF = \mathrm{E}_F\left[D_{CCT}(\theta)\right]^2 \tag{1.19}$$

Donde NCDIF y DTF (*differential test functioning*) son los valores esperados, a través de θ, de la discrepancia cuadrática entre las CCI y entre las CCT, respectivamente. La aportación principal dentro de este marco es la propuesta de un nuevo índice:

$$CDIF_j = \mathrm{E}_F\left(D_{CCI_j}(\theta) \times D_{CCT}(\theta)\right) \tag{1.20}$$

donde CDIF (*compensatory differential item functioning*) es el valor esperado del producto cruzado entre $D_{CCI_j}(\theta)$ y $D_{CCT}(\theta)$. Se cumple que:

$$DTF = \sum_{j=1}^{J} CDIF_j \tag{1.21}$$

donde *J* es el número total de ítems en el banco. Por lo que CDIF es un indicador de en qué grado un ítem contribuye al DTF.

1.5. Métodos de detección del DIF no basados en la TRI

En el presente apartado se describen algunas de las técnicas clásicas que, en este campo, suelen considerarse alternativas a los métodos de TRI y que se han aplicado en el contexto de los tests adaptativos informatizados: el procedimiento de estandarización (Dorans y Kulick, 1983), el procedimiento Mantel-Haenszel (Holland y Thayer, 1988), el modelo de regresión logística (Swaminathan y Rogers, 1990) y el SIBTEST (Shealy y Stout, 1993b). Estos métodos tienen la ventaja de la sencillez de uso y la simplicidad conceptual. Sin embargo, un problema general de todos estos procedimientos es que condicionar a X no es equivalente a condicionar a θ, y por lo tanto, las personas

equiparadas en X pueden diferir en el valor esperado de θ, especialmente en presencia de impacto. Es decir, por lo general:

$$E(\theta \mid X, g = R) \neq E(\theta \mid X, g = F) \tag{1.22}$$

Esto ocurre si X no es un estimador suficiente de θ, especialmente si el test es corto.

1.5.1. Procedimiento de estandarización

Entre las técnicas clásicas, el procedimiento de estandarización destaca por su sencillez conceptual. Puede clasificarse como un método OCI no paramétrico. Se trata del método que tradicionalmente ha utilizado el *Educational Testing Service* para informar del tamaño del DIF de los ítems. La SPD (*standardized p-difference*) se calcula como:

$$SPD = \sum_{k=1}^{K} f_k \left[P_{Fjk} - P_{Rjk} \right] \tag{1.23}$$

donde P_{Fjk} y P_{Rjk} son las probabilidades de acierto para las personas en el grupo de referencia y en el grupo focal que se encuentran en el intervalo k de rasgo. f_k indica la proporción de personas en el intervalo k dentro del grupo focal. SPD es el valor esperado a través de X de la diferencia en la proporción de acierto al ítem entre los grupos comparados (Dorans y Kulick, 1983). Es en sí mismo una medida de tamaño del efecto. Dorans y Holland (1993) proponen la siguiente clasificación para los tamaños del efecto:

A: Ausencia de DIF, el valor absoluto de SPD es menor que 0.05.

B: DIF que indica necesidad de revisión: el valor absoluto de SPD está entre 0.05 y 0.10.

C: DIF que indica necesidad de eliminar el ítem: el valor absoluto de SPD es superior a 0.10.

Dividiendo SPD por su error típico se obtiene un estadístico de contraste z que sigue la distribución normal estándar y permite contrastar la hipótesis nula de ausencia de DIF. Nótese, a partir de la ecuación que SPD, es únicamente apropiado para la detección del DIF unidireccional, ya que en presencia de DIF no unidireccional pueden

cancelarse los efectos del DIF a través de los niveles de rasgo. También se han propuesto medidas "sin signo" para la detección del DIF no unidireccional (Camilli y Shepard, 1994):

$$UPD = \sqrt{\sum_{k=1}^{K} f_k \left(P_{Fjk} - P_{Rjk}\right)^2} \tag{1.24}$$

Que se interpreta igual que *SPD*.

1.5.2. SIBTEST (Simultaneous Item Bias Test)

Shealy y Stout (1993a, 1993b) desarrollaron un modelo OCI no paramétrico para la detección del DIF. En este marco se obtiene como indicador del DIF:

$$\beta_{jSIBTEST}(\theta) = \int \left(P_{jR}(\theta) - P_{jF}(\theta)\right) f(\theta) d\theta \tag{1.25}$$

donde $P_{jR}(\theta)$ y $P_{jF}(\theta)$ marcan la probabilidad de acierto al ítem *j* condicionada al nivel de habilidad (que actúa como variable de igualación) y al grupo de pertenencia del examinado; $f(\theta)$ es la función de densidad de θ ya sea del grupo focal (Bolt, 2000) o del total de la muestra (Nandakumar y Roussos, 2001; Shealy y Stout, 1993a). En realidad, puesto que θ es desconocida se utiliza como estimador la puntuación en el test de anclaje (X_k), y para evitar el sesgo estadístico que se produce por el impacto (ver ecuación 1.22), Shealy y Stout (1993b), propusieron una corrección en la estimación de las proporciones, por lo tanto, el estimador de SIBTEST se calcula como:

$$\hat{\beta}_{jSIBTEST(X_k)} = \sum_{k=0}^{K} f(X_k)\left(P^*_{jR}(X_k) - P^*_{jF}(X_k)\right) \tag{1.26}$$

donde $P^*_{jR}(X_k)$ y $P^*_{jF}(X_k)$ son las proporciones corregidas de acierto en el ítem *j*. Dividiendo $\hat{\beta}_{jSIBTEST(X_k)}$ por su error típico se obtiene un estadístico de contraste *z*, que sigue la distribución normal estándar, y permite comprobar la hipótesis nula de que es cero.

La corrección se establece considerando que *X* es un estimador sesgado de la puntuación verdadera, *V,* y ese sesgo depende de la distribución del rasgo en el grupo al que pertenece la persona. La corrección puede ser lineal (Shealy y Stout, 1993b) o, lo que suele ser más adecuado, no lineal (Jiang y Stout, 1998). En presencia de impacto, las tasas de error Tipo I son extremadamente altas cuando no se utiliza la corrección

(Jiang y Stout, 1998). Esta corrección acerca el SIBTEST a los métodos UCI. Sin embargo, debe tenerse en cuenta que aunque SIBTEST atenúa el problema de la inflación del error Tipo I, este no desaparece y sigue siendo importante en presencia de impacto si el test es corto (Klockars y Lee, 2008; Penfield y Camilli, 2007).

Una ventaja adicional de SIBTEST es que permite determinar si un conjunto de ítems con DIF conduce a un funcionamiento diferencial del test (DTF):

$$\hat{\beta}_{SIBTEST(X_k)} = \sum_{k=1}^{K} f(X_k)\left(E(Y_R^*(X_k) - E(Y_F^*(X_k))\right) \tag{1.27}$$

donde $E(Y_R^*(X_k)$ es el valor esperado de la puntuación en el conjunto de ítems sospechoso. $\hat{\beta}_{SIBTEST(X_k)}$ es una medida sensible al funcionamiento diferencial del test *unidireccional.*

$\hat{\beta}_{jSIBTEST(X_k)}$ puede considerarse una medida de tamaño del efecto análoga a SPD e interpretarse como éste. Roussos y Stout (1996) propone utilizar como criterios de clasificación:

A: DIF nulo o pequeño cuando el valor absoluto de $\hat{\beta}_{jSIBTEST(X_k)}$ es menor que 0.059.

B: DIF que indica necesidad de revisión: el valor absoluto de $\hat{\beta}_{jSIBTEST(X_k)}$ está entre 0.059 y 0.088.

C: DIF que indica necesidad de eliminar el ítem: el valor absoluto de $\hat{\beta}_{jSIBTEST(X_k)}$ es superior a 0.088.

SIBTEST es únicamente apropiado para la detección del DIF unidireccional.

1.5.3. Mantel-Haenszel

El método Mantel-Haenszel (*MH*) es un método que deriva de la investigación biomédica y fue sugerido por Holland (1985; Holland y Thayer, 1988) para la detección del DIF. Puede considerarse un procedimiento OCI no paramétrico. Si no hay DIF, asumiendo que X es un estimador suficiente de θ, debe cumplirse que la razón de ventajas (*odds ratio*) condicional es igual a 1:

$$\alpha_k = \frac{P(u_j = 1 \mid X_k, g = R) / P(u_j = 0 \mid X_k, g = R)}{P(u_j = 1 \mid X_k, g = F) / P(u_j = 0 \mid X_k, g = F)} = 1 \qquad (1.28)$$

Mediante el estadístico χ^2_{MH} (que sigue una distribución χ^2 con un grado de libertad) se contrasta la hipótesis nula de que $\alpha_k = 1$ para cualquier nivel *k* ($\alpha_k = \alpha = 1$). Con este estadístico se contrasta si el número *total* de aciertos en el grupo de referencia coincide con el valor esperado asumiendo que $\alpha = 1$. Además del contraste estadístico, se obtiene una estimación de α, α_{MH} –el cociente de razones común– que es un promedio ponderado de los distintos α_k. Para facilitar su interpretación se calcula el logaritmo, $Ln(\alpha_{MH})$. Un valor $Ln(\alpha_{MH}) = 0$ indica ausencia de sesgo; valores positivos implican que el ítem favorece al grupo de referencia y valores negativos que favorece al grupo focal. Si los ítems siguen el modelo de Rasch y α_k es igual para a través de los niveles *k*, $Ln(\alpha_{MH})$ es una estimación de la diferencia entre los parámetros *b* (Zwick, Thayer y Wingersky, 1993):

$$Ln(\alpha_{MH}) = b_F - b_R \qquad (1.29)$$

Si los datos se ajustan al modelo de dos parámetros y las CCI no difieren en pendiente a través de los grupos, $Ln(\alpha_{MH})$ es un estimador de (Roussos, Schnipke y Pashley, 1999):

$$Ln(\alpha_{MH}) = a(b_F - b_R) \qquad (1.30)$$

En el caso del modelo de tres parámetros no existe una correspondencia simple entre $Ln(\alpha_{MH})$ y los parámetros de los ítems, ya que la *odss ratio* no es constante a través de θ (Roussos et al., 1999).

El valor de $Ln(\alpha_{MH})$ puede entenderse como una medida de tamaño del efecto. El ETS usa tres categorías para reflejar el grado de DIF en el ítem (Zieky, 1993)[1]:

[1] En realidad el ETS trabaja en la métrica delta, en la que se multiplican por –2.35 $Ln(\alpha_{MH})$ y los puntos de corte para clasificar el tamaño del DIF.

A: DIF nulo o pequeño cuando el valor absoluto de $Ln(\alpha_{MH})$ es estadísticamente distinto de 0 y menor que 0.43.

B: DIF que indica necesidad de revisión: el valor absoluto de $Ln(\alpha_{MH})$ no es estadísticamente superior a 0.43 o está entre 0.43 y 0.64.

C: DIF que indica necesidad de eliminar el ítem: el valor absoluto de $Ln(\alpha_{MH})$ es estadísticamente superior a 0.43 y superior a 0.64.

$Ln(\alpha_{MH})$ se encuentra en una métrica distinta a $\hat{\beta}_{jSIBTEST(X_k)}$. Sin embargo, a pesar de que son muy diferentes conceptualmente, existe una fuerte relación entre $\hat{\beta}_{jSIBTEST(X_k)}$ y $Ln(\alpha_{MH})$. Dorans y Holland (1993) han mostrado que, en estudios de simulación, ambos indicadores correlacionan más de 0.95. Siguiendo a Shealy y Stout (1993b) la relación puede ser aproximada como (citado en Bolt y Stout, 1996):

$$\hat{\beta}_{jSIBTEST(X_k)} = \frac{2.35 Ln(\alpha_{MH})}{15} \tag{1.31}$$

Así pues, los puntos de corte sugeridos para $Ln(\alpha_{MH})$ se corresponderían con unos puntos de corte para SIBTEST de 0.067 y 0.10.

Nótese que α_{MH} es un promedio ponderado de los distintos α_k. Por tanto, únicamente es apropiado para la detección del DIF unidireccional, ya que en presencia de DIF no unidireccional pueden cancelarse los efectos del DIF a través de los niveles de rasgo.

1.5.4. Regresión logística

Swaminathan y Rogers (1990) proponen el uso de la regresión logística a la detección del DIF. Se trata de un procedimiento OCI paramétrico. En este modelo, el logit de la probabilidad de acierto puede ser función lineal de la puntuación en el test (X_k), la pertenencia al grupo (g, puntuado como 0 o 1) y la interacción entre ambos ($X_k \times g$):

$$\ln\left(\frac{P(u_{ij}=1 \mid X_k)}{P(u_{ij}=0 \mid X_k)}\right) = \beta_0 + \beta_1 X_k + \beta_2 g + \beta_3 X_k g \tag{1.32}$$

Donde β_0 corresponde al intercepto del modelo (o línea base) y β_1 determina en qué medida la probabilidad de acertar depende del nivel de rasgo. Pueden contrastarse comparando modelos anidados, partiendo del modelo más general que incluye todos los parámetros. Si no puede mantenerse que $\beta_3 = 0$, hay DIF *no unidireccional*. Si manteniendo $\beta_3 = 0$ no puede mantenerse que $\beta_2 = 0$, hay DIF *unidireccional*. Si puede mantenerse $\beta_2 = \beta_3 = 0$, puede mantenerse que no hay DIF.

Para medir el tamaño del efecto del DIF se ha propuesto acompañar las medidas de significación estadística con los valores de incremento en R^2 entre modelos anidados pero no existe acuerdo en los puntos de corte (Jodoin y Gierl; 2001; Zumbo y Thomas, 1997).

Alternativamente, otra medida de tamaño del efecto para el DIF unidireccional es el peso de regresión β_2 (Monahan, McHorney, Stump y Perkins, 2007) Asumiendo el modelo de Rasch, β_2 puede considerarse un estimador de la diferencia entre los parámetros *b:*

$$\beta_2 = b_F - b_R \tag{1.33}$$

y se interpretaría como $Ln(\alpha_{MH})$. Una ventaja del procedimiento de regresión logística es que, frente a otras técnicas clásicas como Mantel-Haenszel, permite detectar el DIF no unidireccional. Sin embargo, se asume que: (a) *X* es un estimador adecuado de θ, y (b) la relación entre la puntuación en el ítem y θ sigue el modelo 2PL.

1.6. Resultados de comparación de los métodos

Es difícil establecer cuál de los métodos descritos es mejor en términos de control de las tasas de error Tipo I y tamaño de la potencia. Cada método tiene ventajas e inconvenientes en función de características como la longitud del test, la presencia de impacto, el tamaño de las muestras, el tipo de DIF o el ajuste a un modelo paramétrico, entre otras (Gómez-Benito, Hidalgo y Guilera, 2010).

Welkenhuysen-Gybels (2004) considera que las técnicas UCI son preferibles teóricamente a los procedimientos OCI ya que se trabaja directamente al nivel de la variable latente. Por ejemplo, los procedimientos de TRI tienen la ventaja (frente a otras alternativas) de que las CCI son directamente comparadas. En las técnicas tradicionales

se empareja a los evaluados por su puntuación en el test y las tasas de error Tipo I pueden inflarse cuando el test es poco fiable (p.e., tests cortos), los ítems varían en discriminación y las distribuciones del nivel de rasgo difieren a través de los grupos (Monahan y Ankenman, 2010). Los evaluados emparejados por su puntuación observada no están necesariamente emparejados en el rasgo latente. Por ejemplo, DeMars (2010) muestra que cuando hay impacto la diferencia en θ entre el grupo de referencia y el grupo focal para grupos de evaluados emparejados en la puntuación observada se incrementa a medida que la fiabilidad decrece. Este problema es mayor en muestras grandes y se atenúa a medida que se incrementa la longitud del test (Bolt y Gierl, 2006; DeMars, 2009; Jiang y Stout, 1998; Li y Stout, 1996). Pei y Li (2010) también encuentran que diferencias en la variabilidad del nivel de rasgo entre grupos inflan el error Tipo I de SIBTEST, regresión logística y Mantel-Haenszel, pero no de IRT LR test.

Una posible solución a este problema es aplicar las correcciones propuestas para SIBTEST a los procedimientos Mantel-Haenszel y de Regresión Logística. Esto lleva a un mejor control en la tasa de error Tipo I, sobre todo en las condiciones más adversas: presencia de impacto y muestras grandes (DeMars, 2009). No obstante si el test es demasiado corto (p.ej., 10 ítems) las tasas de error Tipo I pueden seguir estando infladas (Klockars y Lee, 2008).

Otra posibilidad es sustituir la variable de igualación X por un estimador más apropiado dadas las características de los datos. Por ejemplo, Monahan y Ankenman (2010) utilizan $\hat{\theta}_{3PL}$ como variable de igualación en Mantel-Haenszel y encuentran que se mantienen tasas de error Tipo I apropiadas, sin necesidad de establecer ninguna corrección adicional. Una limitación de esta aproximación es que supondría compartir las limitaciones de los métodos UCI.

En relación al tipo de DIF, varias de las técnicas clásicas (estandarización, MH, SIBTEST) comparten la limitación de no ser adecuadas para la detección del DIF no unidireccional. Aunque se han propuesto adaptaciones para detectar el DIF no unidireccional (p.e., Li y Stout, 1996; Mazor, Clauser y Hambleton, 1994), la TRI y la regresión logística proporcionan soluciones más directas. Así, si el DIF es no unidireccional la regresión logística muestra mayor potencia que la aproximación Mantel-Haenszel (Hidalgo y López-Pina, 2004; Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990). Sin embargo, la TRI puede ser una aproximación más

adecuada pues la regresión logística falla cuando los datos no se ajustan al modelo de dos parámetros. DeMars (2009) encuentra que, en presencia de impacto, se produce inflación de la tasa de error Tipo I para la regresión Logística en la detección del DIF no unidireccional, independientemente de que se aplique la corrección.

En relación al tamaño de la muestra, ya hemos observado que afecta a la inflación del error tipo I de varias técnicas clásicas. Sin embargo, cuando las condiciones son adecuadas (i.e., DIF unidireccional, ausencia de impacto, etc.), los métodos clásicos obtienen su mayor ventaja. Mantel-Haenzsel, estandarización y SIBTEST muestra tasas de error Tipo I adecuadas y tasas de potencia similares incluso en muestras pequeñas (Nandakumar, 1993; Narayanan y Swaminathan, 1994; Roussos y Stout, 1996). La regresión logística también muestra buen rendimiento aunque, en presencia de DIF unidireccional, Mantel-Haenszel tiene mayor potencia que la regresión logística (Hidalgo y López-Pina, 2004).

Por el contrario, los métodos de TRI requieren muestras mayores para una estimación adecuada de los parámetros y la comprobación del ajuste de los datos al modelo paramétrico. Esto último, la comprobación del ajuste de los datos al modelo, es probablemente el punto más desfavorable para el uso de las técnicas de TRI. Tal comprobación requiere muestras grandes y, en el caso de que las respuestas no se ajusten a un modelo al uso (p.e., 3PL), el ítem no debería ser incluido en los análisis. Cuando los supuestos son correctos, las tasas de error Tipo I para el IRT LR test son próximas a su valor nominal y la potencia estadística se incrementa con el tamaño muestral, la discriminación del ítem estudiado y el tamaño del DIF (e.g., Ankenmann, Witt y Dunbar, 1999; Lopez-Rivas, Stark y Chernyshenko, 2008; Stark, Chernyshenko y Drasgow, 2006) pero si los supuestos son incorrectos, las tasas de error Tipo I pueden estar infladas (Bolt, 2002) y se incrementarán a medida que el tamaño de la muestra aumenta. Por ejemplo, Woods (2008b) muestra que cuando las distribuciones latentes no siguen la distribución normal para uno o ambos de los grupos, las tasas de error Tipo I estarán infladas, los parámetros no estarán bien estimados y las diferencias entre los grupos en la media y la varianza de la distribución del rasgo latente estarán sobreestimadas. Los métodos no paramétricos como MH, la estandarización o SIBTEST no tienen este problema pues no asumen un modelo concreto para el ítem.

Capítulo 2

Tests Adaptativos Informatizados

El presente capítulo tiene dos objetivos. El primero es presentar la estructura y los procedimientos fundamentales utilizados en los tests adaptativos informatizados (TAIs). El segundo objetivo es describir algunas particularidades en el control y estudio de las propiedades psicométricas en un TAI.

2.1. Tests adaptativos informatizados

Los TAIs resultan posibles gracias al desarrollo en dos áreas: la psicométrica, principalmente por los desarrollos de la teoría de respuesta al ítem (TRI), y la informática, con un abaratamiento de equipos y mayor capacidad de cálculo. Se denominan test adaptativos porque el nivel de dificultad de los ítems presentados se va adaptando al nivel de habilidad que se va estimando al evaluando a lo largo del test; esto es, la selección de ítems se realiza tomando en cuenta la respuesta a los ítems previos. En términos generales, si un evaluado responde de manera correcta a un ítem, el siguiente ítem administrado es más difícil; si responde de manera incorrecta, se presenta un ítem con menor grado de dificultad. La estimación final del nivel de rasgo de los evaluados no es un recuento de ítems acertados, sino que se considera tanto el patrón de aciertos y fallos como el vector de parámetros de los ítems recibidos. Por tanto, ni todos los evaluados han de recibir el mismo conjunto de ítems ni igual proporción de aciertos conlleva igual nivel de rasgo estimado. El que dos personas puedan ser comparadas en el nivel de rasgo, a pesar de no haber respondido a los mismos ítems, resulta posible gracias a las propiedades de invarianza en las estimaciones e independencia local de los ítems de la TRI (Wainer, 2000a).

Los TAIs ofrecen diversas ventajas en comparación con los test convencionales de lápiz y papel. Por ejemplo, en los TAIs es posible (Olea y Ponsoda, 2003):

- Tener una precisión en la estimación similar para los diferentes niveles de habilidad.
- Mantener la precisión en la estimación con un número menor de ítems.
- Incorporar procedimientos que preserven la seguridad del test.
- Aplicar ítems en formatos innovadores.

Estas ventajas hacen de los TAIs instrumentos especialmente oportunos para evaluaciones masivas o continuadas como en la evaluación educativa, o en situaciones donde es fundamental conocer el nivel de habilidad de manera precisa, como en contextos de selección de personal o cuando se otorgan certificaciones para acreditar el ejercicio de alguna profesión.

Diversos test con amplia tradición se aplican actualmente de manera adaptativa como el LSAT (*Law School Admission Test*), el ASVAB (*Armed Services Vocational Aptitude Battery*), el TOEFL (*Test of English as a Foreign Languaje*) o el GMAT (*Graduate Management Adminission Test*). En España existen al menos tres test adaptativos informatizados que se comercializan: (a) el TRASI, que es un test de razonamiento analítico, secuencial e inductivo (Rubio y Santacreu, 2003); (b) eCAT, un test para la evaluación del conocimiento del inglés, tanto de gramática (eCAT-Grammar; Olea, Abad, Ponsoda y Ximénez, 2004) como de compresión oral (eCAT-Listening; Olea, Abad, Ponsoda, Barrada, y Aguado, en prensa); y (c) el CAT-Health que evalúa calidad de vida (Rebollo, García-Cueto, Zardaín, Cuervo, Martínez, Alonso, Ferrer y Muñiz, 2009). En México, el test EXHCOBA mide habilidades y conocimientos básicos y se emplea para el ingreso a la universidad (Backhoff, Ibarra, Rosas y Larrazolo, 1999).

A pesar de las ventajas que ofrece un TAI, el proceso para desarrollarlo hasta que se encuentra operativo para su aplicación es muy costoso y exige una importante inversión de trabajo por parte de expertos en las áreas de contenido que mide el test, en psicometría y en informática. Wainer (2000a) recomienda que el uso de TAIs se limite a

los casos en los que la aplicación informatizada permita medir mejor el rasgo y cuando el test se aplique de manera continuada.

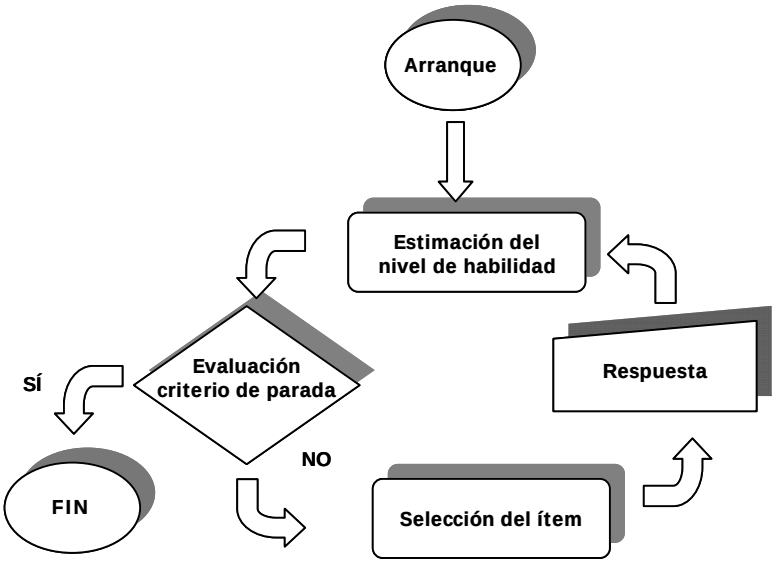
2.2. Componentes fundamentales de un TAI

En general, la aplicación de un TAI requiere de un banco de ítems y de un algoritmo adaptativo. En el banco de ítems se definen los contenidos evaluados, con las especificaciones que se consideren oportunas. El banco de ítems se puede considerar como el elemento estático del sistema y es una base de datos que contiene los ítems, la información psicométrica de dichos ítems, así como información relevante para su aplicación (p.ej., enunciado, estatus –operativo, ítem pretest, ítem retirado...–, categorías de contenido que evalúa, etc.).

El banco de ítems es un elemento imprescindible en un TAI y un algoritmo adaptativo puede funcionar mejor cuanto mayor sea la calidad del banco en términos de representación de contenidos, tamaño, fiabilidad y validez (Flaugher, 2000). Son las fases de construcción y el mantenimiento del banco las que permiten establecer la calidad del banco y obtener la información necesaria para aplicar el algoritmo adaptativo.

El algoritmo del TAI define el funcionamiento del mismo. Los procesos implicados en el algoritmo adaptativo se pueden organizar en un diagrama de flujo (ver Figura 2.1).

Figura 2.1. Diagrama de flujo de un TAI (tomado de Barrada, en prensa).



Como se muestra en la Figura 2.1, el primer proceso es el *Arranque.* En esta fase se inicia el sistema, se incorpora la información contenida en el banco de ítems y los criterios no definidos como constantes en el algoritmo. Se presentan las instrucciones y se piden los datos de la persona. Posteriormente se establece la secuencia *Estimación del nivel de habilidad* → *Evaluación del criterio de parada* → *Selección y aplicación del ítem* → *Registro de la respuesta del evaluado*, hasta que se cumple el criterio de parada y el TAI finaliza. En los siguientes apartados se describen brevemente las distintas posibilidades en relación a cada uno de esos procesos.

2.3. Estimación del nivel de habilidad

El objetivo principal de ésta fase es estimar de forma precisa el nivel de habilidad del evaluado. En un primer momento, cuando el evaluado no ha respondido a ningún ítem, se le asigna un nivel de rasgo provisional que puede ser la media poblacional del nivel de habilidad o, para evitar que se presente siempre el mismo ítem al inicio de la evaluación, un nivel de habilidad aleatorio en un rango de valores (p.ej., un valor entre –0.4 y 0.4, si la media es cero). Otras posibilidades, menos frecuentes, son considerar el uso de información previa (van der Linden, 1999) o permitir que el evaluado elija el nivel de dificultad del ítem.

Por otro lado, cuando el evaluando ya ha respondido a algunos ítems las alternativas más frecuentes para la estimación son máxima verosimilitud (ML) o el uso de un procedimiento bayesiano de estimación. Con ML a cada evaluado *i* se le asigna como nivel de rasgo, $\hat{\theta}_{iq}$, aquel que maximiza la función de verosimilitud de las respuestas a los *q* ítems aplicados ($\mathbf{u}_{i(obs)}$):

$$f\left(\mathbf{u}_{i(obs)} \mid \theta, \mathbf{m}_i\right) = \prod_{j=1}^{J} P_j(u_{ij} \mid \theta_i)^{m_{ij}} \tag{2.1}$$

donde los elementos de $\mathbf{u}_i$ del evaluado *i*, pueden ser $u_{ij} = 1$, si la respuesta al ítem *j* es correcta, $u_{ij} = 0$ si es incorrecta y cualquier otro número (p.ej., $u_{ij} = 9$) si dicho ítem no se ha aplicado. $\mathbf{m}_i$ es el vector que contiene a los elementos m_{ij} que indican si el ítem *j* se ha aplicado ($m_{ij} = 1$) o no se ha aplicado ($m_{ij} = 0$). $P_j(u_{ij} \mid \theta_i)$ denota la probabilidad de respuesta correcta al ítem *j* de un banco con *J* ítems. Un problema del método ML es que no permite obtener estimaciones finitas de θ mientras una persona tenga un patrón constante de respuestas.

En los métodos de estimación bayesiana se incorpora información sobre la distribución previa de la habilidad $f(\theta \mid \boldsymbol{\pi})$ a la función de verosimilitud, para obtener la distribución posterior que se calcula como:

$$f\left(\theta \mid \mathbf{u}_{i(obs)}, \mathbf{m}_i, \boldsymbol{\pi}\right) = \frac{f\left(\mathbf{u}_{i(obs)} \mid \theta, \mathbf{m}_i\right) f(\theta \mid \boldsymbol{\pi})}{\int f\left(\mathbf{u}_{i(obs)} \mid \theta, \mathbf{m}_i\right) f(\theta \mid \boldsymbol{\pi}) d(\theta)} \tag{2.2}$$

donde $\boldsymbol{\pi}$ es el vector de parámetros que determinan la distribución de la habilidad en la población θ. Dos métodos pueden ser utilizados para estimar $\hat{\theta}_{i(q)}$: asignar la moda de la distribución posterior (método MAP o Máximo a posteriori) o la media de la distribución posterior (método EAP o Esperanza a posteriori). Los métodos de estimación bayesiana permiten la estimación con patrones constantes. Una ventaja de EAP es que matemáticamente es muy fácil de implementar.

En general los métodos bayesianos producen menor error típico pero mayor sesgo que ML, ya que el nivel de habilidad del evaluado depende de la distribución del rasgo en la población, $f(\theta \mid \boldsymbol{\pi})$, y, las estimaciones están sesgadas hacia la media cuando el test es corto, lo que puede favorecer a los sujetos de bajo nivel de habilidad (Wang y Vispoel, 1998). Con ML se sobrestiman los niveles alto de rasgo y se subestiman los bajos, pero en un TAI el sesgo con ML suele ser menor (Abad, Olea,

Real y Ponsoda, 2002). En cualquier caso, las diferencias entre los métodos de estimación suelen ser pequeñas cuando el TAI es suficientemente largo (p.ej., 30 ítems).

Al problema de los patrones constantes en la estimación ML se le han dado varias soluciones, entre ellas:

a) El método de Dodd (1990) propone, si el patrón es de aciertos, asignar como nivel de habilidad

$$\hat{\theta}_{iq} = \frac{\hat{\theta}_{i(q-1)} + b_{\max}}{2}, \tag{2.3}$$

mientras que si el patrón es de errores el nivel asignado se calcula como

$$\hat{\theta}_{iq} = \frac{b_{\min} - \hat{\theta}_{i(q-1)}}{2}, \tag{2.4}$$

donde $b_{\max}$ y $b_{\min}$ son, respectivamente, los parámetros de dificultad del ítem más fácil y más difícil del banco.

b) Calcular el valor esperado de θ, considerando una función de densidad normal truncada [$f(\theta)$]; el truncamiento se realiza entre $\hat{\theta}_{i(q-1)}$ y $b_{\max}$ cuando se acierta y entre $b_{\min}$ y $\hat{\theta}_{i(q-1)}$ cuando se falla. Se ha comprobado que mediante este procedimiento, el valor de la función de información proporcionada por el segundo ítem se mantiene en niveles aceptables (Revuelta y Ponsoda, 1997).

c) Aplicar directamente métodos de estimación bayesiana como MAP o EAP (Wainer y Mislevy, 2000). En realidad, tanto el método de Dodd como la propuesta de Revuelta y Ponsoda (1997) se pueden justificar desde el punto de vista bayesiano. En ambos procedimientos se asume una distribución previa truncada ya sea uniforme, en el método de Dodd, o normal, en el método de Revuelta y Ponsoda (Olea, Ponsoda, Revuelta, Hontangas y Suero, 1999).

2.4. Evaluación del criterio de parada

Estos criterios se establecen a partir de los objetivos de la aplicación y las propiedades psicométricas del banco (Olea y Ponsoda, 2003). Los criterios más utilizados son: (a) un criterio de longitud fija, que consiste en detener el test cuando se han aplicado un número predefinido de ítems; (b) un criterio de longitud variable, que

detiene la aplicación cuando el error típico de estimación de $\hat{\theta}_{iq}$ es menor que un criterio predefinido; o c) un criterio mixto, que resulta de la combinación de los dos anteriores.

El error típico de estimación de aproximación de ML, $\sigma_{\hat{\theta}|\theta}$, se calcula como:

$$\sigma_{(\hat{\theta}|\theta)} = \frac{1}{\sqrt{I(\hat{\theta})}} \tag{2.5}$$

donde $I(\hat{\theta})$ es la función información del test (IT), que se calcula como:

$$TI(\theta) = \sum_{j=1}^{J} I_j(\theta) m_{ij} \tag{2.6}$$

En el caso del modelo de tres parámetros:

$$I_j(\theta) = D^2 \frac{a_j{}^2 [P_j(u_{ij} = 0 \mid \theta)][P_j(u_{ij} = 1 \mid \theta) - c_j)]^2}{P_j(u_{ij} = 1 \mid \theta)(1 - c_j)^2} \tag{2.7}$$

La función de información tiene dos características. Por un lado, que es aditiva, por lo que cada ítem va añadiendo información. Por otro, que no depende de los valores de las respuestas a los ítems. Esto significa que la información de un test para un valor de habilidad concreto es el resultado de sumar las informaciones de todos los ítems que componen el test (Wainer y Mislevy, 2000).

En relación a qué criterio de parada es más adecuado:

- Los *test de longitud fija* suelen ser adecuados cuando el objetivo es obtener una primera evaluación, por ejemplo cuando se quiere conocer el nivel de conocimientos para el acceso a determinado nivel educativo. La desventaja de este criterio es que los niveles de habilidad se miden con diferentes niveles de precisión, sobre todo en los niveles de habilidad extremos, ya que los bancos difícilmente tienen ítems igualmente informativos para todos los niveles de habilidad.
- Los *test de longitud variable* se eligen cuando las consecuencias de la evaluación son especialmente importantes y se desea conocer el nivel de habilidad con precisión. Por ejemplo, en los procesos de selección de personal o

cuando se otorgan certificaciones, como el Law School Admission Test. La desventaja de este criterio es de validez aparente y se relaciona con el hecho de que los evaluados pueden percibir que quienes respondieron a más ítems tuvieron más oportunidad de demostrar sus habilidades.

- Los *test con criterio de parada mixto* suelen utilizarse en los test referidos al criterio. En este caso, el test termina cuando se han aplicado un número determinado de ítems, siempre y cuando el nivel de habilidad esté fuera del intervalo de confianza de la puntuación de corte que permite clasificar a los evaluados (en apto/no apto o aprobado/suspenso, etc.). De lo contrario, el test continúa hasta que se pueda determinar con un error mínimo si están por arriba o por debajo de dicho punto de corte (Olea y Ponsoda, 2003).

Si se cumple el criterio de parada, el test finaliza. Se archivan las respuestas, se actualizan las variables que dependen de las aplicaciones (p.e., tasas de exposición de los ítems) y se genera un informe automatizado para el usuario en el que se especifique de forma clara su rendimiento.

2.5. Criterios de selección de ítems

La selección de ítems es la parte medular del algoritmo adaptativo y debe cumplir diversos objetivos tales como (Davey y Parshall, 1995; Stocking y Lewis, 2000):

- Establecer estrategias de selección de ítems que maximicen *la eficiencia del test.*
- Garantizar que la selección se ajuste a las *especificaciones de contenido*.
- Mejorar la *seguridad* del test controlando la exposición de los ítems sobre-expuestos.
- Gestionar adecuadamente el banco de ítems aumentando la exposición de los ítems infrautilizados.

Esta fase se considera la más compleja, porque los procedimientos que se utilizan para cumplir alguno de los objetivos pueden comprometer el cumplimiento de los sucesivos. Por ejemplo, el criterio que selecciona el ítem más informativo tiende a

disminuir la eficacia en la gestión del banco, porque se selecciona un porcentaje pequeño de ítems del banco. De igual forma, si se desea asegurar que el test mida múltiples contenidos se altera la eficiencia en la selección por máxima información. Por lo tanto, los procedimientos creados para cumplir alguno de los objetivos deben adecuarse para interferir lo menos posible en el cumplimiento de los demás propósitos.

Dentro de los procedimientos que se presentan a continuación están: (a) los algoritmos para *la selección eficiente de ítems*; (b) los métodos para introducir *restricciones en el contenido del test*; y (c) procedimientos de *control de la exposición*. Siguiendo la propuesta de Barrada (en prensa) distinguiremos entre los métodos que evitan la selección de los ítems sobre-expuestos y los procedimientos que permiten aumentar la exposición de los ítems infra-utilizados, lo cual está relacionado con la seguridad en el test y con una mejor gestión del banco de ítems.

2.5.1. Estrategias de selección de ítems

Sea $h_{i,q}$ el identificador del ítem seleccionado para el examinado *i* y la posición *q* en el test. Por tanto, $h_{i,q}$ es un número natural entre 1 y *J*. La estrategia de selección de ítems más comúnmente utilizada es el *método de máxima información*. Este método consiste en seleccionar el ítem que hace máxima la información para el nivel provisional de habilidad estimado ($\hat{\theta}_{i(q)}$) y se escribe como:

$$h_{i(q+1)} = \arg\max_{j \in B_{i(q+1)}} I_j(\hat{\theta}_{iq}) \tag{2.8}$$

donde $B_{i(q+1)}$ define aquellos ítems del banco que pueden ser seleccionados. Esta estrategia de selección de ítems puede presentar limitaciones en dos de los objetivos a optimizar en los TAIs: la precisión y la seguridad del test.

En cualquier caso, el problema asociado *a estos procedimientos que se basan sólo en la precisión* tienen que ver con que se selecciona el ítem más informativo para un nivel de habilidad provisional sin tomar en cuenta el error de estimación. Si éste es grande el nivel de habilidad verdadero va a estar alejado del nivel de habilidad para el cual se ha seleccionado el ítem, lo que ocurre especialmente cuando son pocos los ítems aplicados. Algunas alternativas que se han propuesto para resolver este problema son:
Método de máxima información ponderada por verosimilitud: Para tomar en cuenta el error de medida, Veerkamp y Berger (1997) propusieron ponderar la función de

información por la función de verosimilitud. De modo que el ítem se selecciona mediante la siguiente función:

$$h_{i(q+1)} = \arg\max_{j \in B_{i(q+1)}} \int I_j(\theta) f\left(\mathbf{u}_{i(obs)} \mid \theta, \mathbf{m}_i\right) d\theta \tag{2.9}$$

Así, cuando se han aplicado pocos ítems el error de estimación es mayor y la función de verosimilitud tiende a ser más plana, por lo que en vez de buscar el valor máximo en el nivel estimado (puntual), se busca la información para un rango mayor de niveles de habilidad.

- *Método basado en la información de Kullback-Leibler*: Una función alternativa a la función de información de Fisher es la función de información de Kullback-Leibler, $KL_j(\hat{\theta}_{iq})$, propuesta por Chang y Ying (1996) como regla de selección en los TAIs. $I_j(\hat{\theta}_{iq})$ es una estimación de la capacidad de discriminación de un ítem entre el nivel de habilidad verdadero (que se estima por $\hat{\theta}_{iq}$) y niveles de habilidad *próximos a* $\hat{\theta}_{iq}$. El uso de esta medida de información puede resultar inapropiado al principio del test, cuando $\hat{\theta}_{iq}$ está alejada del valor verdadero. $KL_j(\hat{\theta}_{iq})$ es una estimación de la capacidad de discriminación de un ítem entre el nivel de habilidad verdadero con respecto a un conjunto de niveles de habilidad *plausibles* dado el número de ítems aplicados. Una medida de discrepancia en la distribución de la respuesta al ítem *j,* según sea función de $\hat{\theta}_{iq}$ o θ, es:

$$KL_j(\theta \,||\, \hat{\theta}_{iq}) = P_j(u_{ij}=1 \mid \hat{\theta}_{iq}) \log\left[\frac{P_j(u_{ij}=1 \mid \hat{\theta}_{iq})}{P_j(u_{ij}=1 \mid \theta)}\right] + P_j(u_{ij}=0 \mid \hat{\theta}_{iq)}) \log\left[\frac{P_j(u_{ij}=0 \mid \hat{\theta}_{iq})}{P_j(u_{ij}=0 \mid \theta)}\right]$$

(2.10)

El ítem a elegir será el que tengo el mayor valor de $KL_j(\hat{\theta}_q)$, en el que se considera (θ_1, θ_u) como un intervalo en torno al último nivel de habilidad estimado ($\hat{\theta}_{iq}$) para definir qué θ son plausibles:

$$h_{i(q+1)} = \arg\max_{j \in B_{i(q+1)}} \int_{\theta_1}^{\theta_u} KL_j(\theta \,||\, \hat{\theta}_{iq}) d\theta \tag{2.11}$$

La amplitud del intervalo *se va reduciendo* a medida que aumenta el número de ítems.

- *Método basado en la información de Kullback-Leibler ponderando por la función de verosimilitud:* Respecto de la función anterior, Chang y Ying (2006) propusieron ponderarla por la función de verosimilitud, por lo que el ítem elegido es aquel cuyo valor sea máximo en:

$$h_{i(q+1)} = \arg\max_{j \in B_{i(q+1)}} \int KL_j(\theta \,\|\, \bar{\theta}_{iq}) f\left(\mathbf{u}_{i(obs)} \mid \theta, \mathbf{m}_i\right) d\theta \tag{2.12}$$

Los estudios que comparan diversos criterios de selección señalan que las diferencias en precisión por el uso de diferentes reglas de selección desaparecen a medida que aumenta el número de ítems administrados. Sin embargo, la tasa de convergencia de los métodos puede depender de las propiedades psicométricas del banco. Por ejemplo, Chen, Ankermann y Chang (2000) encuentran que después de aplicar 10 ítems las diferencias entre métodos desaparecen, mientras que en Barrada, Olea, Ponsoda y Abad (2009) las diferencias persisten con longitudes mayores.

2.5.2. Especificaciones de contenido en los test

Es común que en los programas de evaluación se desee controlar diversas especificaciones que no necesariamente están relacionadas con criterios estadísticos. Por ejemplo, los bancos de ítems aun cuando midan una dimensión dominante, pueden tener ítems que pertenezcan a diversos subdominios. Si la selección de ítems se basa únicamente en criterios como el de máxima información, el resultado puede ser un test con contenidos no balanceados. Para evitar esta y otras situaciones se construye una tabla de especificaciones en las que pueden señalarse (Ponsoda, Revuelta, Hontangas y Suero, 1999):

- Restricciones basadas en las propiedades intrínsecas de los ítems (proporción de ítems por contenido, formato de los ítems, posición de la respuesta correcta).
- Restricciones en la presentación de ítems con contenido similar. O por el contrario, ítems que deben aparecer juntos (si todos ellos se refieren a un mismo estímulo como en el caso de los testlets).

- Restricciones en relación a ítems que van dirigidos exclusivamente a poblaciones específicas, etc.

Formalmente, cada especificación impone una restricción en la selección de ítems en el banco. Como consecuencia, con el algoritmo del TAI se debe maximizar la información estadística, manteniendo el cumplimiento de las especificaciones no estadísticas (van der Linden, 2000). Los siguientes métodos se han planteado para introducir dichas especificaciones al algoritmo.

- *Dividir el banco de ítems:* Esta propuesta de Kingsbury y Zara (1991) consiste en dividir el banco de acuerdo a los atributos de los ítems. Para la selección se propone una adaptación del criterio de máxima información en la que, para mantener el balance de contenidos del test, la selección del ítem se realiza dentro de la categoría con una proporción menor de ítems aplicados con respecto al número de ítems a aplicar.
- *Método de las desviaciones ponderadas:* En el modelo propuesto por Stocking y Swanson (1993) los expertos en áreas de contenido establezcan las restricciones necesarias y les asignen un peso (w_g) dependiendo su importancia. Por otro lado, cada ítem en la base de datos tiene señalado el nivel de cumplimiento de dicha restricción. La precisión en la medida se considera un objetivo más con su correspondiente peso. En este sentido, los límites especificados en el diseño del test dejan de ser considerados como objetivos estrictos y pasan a ser objetivos deseables. Los ítems seleccionados son aquellos que minimizan la suma de las desviaciones ponderadas entre los objetivos del test y lo que se obtendría si se aplicara dicho ítem:

$$h_{i(q+1)} = \arg\min_{j \in B_{i(q+1)}} \sum_{g=1}^{G} w_g \left| \pi_{jg} - \gamma_g \right| \tag{2.13}$$

  donde *G* es el número de restricciones incluidas en el TAI, π_{ig} es el valor para la restricción *g* en el caso de ser aplicado el ítem j y γ_g es el objetivo para la restricción *g*.
- *Tests sombra:* van der Linden (2000) propone el uso de la programación lineal para el control de las especificaciones del TAI. Para seleccionar cada ítem, se construye primero un test completo que incluye los ítems aplicados y satisface

todas las restricciones, siendo óptimo según las reglas de selección. Posteriormente se selecciona como ítem a aplicar el ítem que resulta óptimo según la regla de selección.

El método de Kingsbury y Zara (1991) es bastante simple, pero tiene la desventaja de que se puede aplicar en pocas situaciones (Barrada, en prensa). El método de las desviaciones ponderadas presenta dos problemas: (a) no se garantiza un cumplimiento estricto de las restricciones; y (b) el cumplimiento de las restricciones se va deteriorando a medida que avanza el test (Cheng y Chang, 2009; van der Linden, 2000, 2005). El test en la sombra parece ser el más eficiente en todos los sentidos, pero su complejidad matemática e informática restringe su uso (Diao y van der Linden, en prensa).

2.5.3. Control de la exposición y seguridad del test

Preservar el banco de ítems es un aspecto importante en algunas situaciones de evaluación. La importancia de limitar el conocimiento previo de los ítems en los contextos de evaluación depende del uso de las puntuaciones obtenidas en el test. Si éstas son utilizadas para tomar decisiones que tienen consecuencias para los evaluados, como la admisión a la universidad o a un puesto de trabajo, entonces se deben invertir recursos que limiten la probabilidad de filtración de los ítems.

Los criterios de selección de ítems que únicamente toman en cuenta la eficiencia en la estimación del nivel de habilidad favorecen la selección de ítems con ciertas propiedades métricas (Stocking y Lewis, 2000). Por ejemplo, en el modelo logístico de 3 parámetros la selección de ítems por máxima información lleva a la selección de ítems con mayor parámetro *a* y menor parámetro *c*.

Las reglas de selección de ítems que buscan maximizar la eficacia en la estimación tienden a presentar los mismos ítems a diferentes examinados por dos motivos: (a) al comienzo del test hay poca variabilidad en los niveles de habilidad estimados; y (b) los ítems seleccionados se concentran entre aquellos con mayor parámetro de discriminación (Li y Schafer, 2005). Por ello, muchos evaluados comparten un porcentaje muy alto de ítems. El posible efecto adverso para la seguridad se agrava puesto que en los TAIs los bancos de ítems son relativamente estáticos a lo largo del tiempo y la aplicación se hace de manera continuada, lo cual supone que las

personas puedan recurrir a otras personas, academias de preparación o a sitios web especializados en el filtrado de preguntas (Chang, 2004; Davey y Nering, 2002).

Dos son las variables empleadas como medida indirecta del efecto de la filtración de ítems (Chang y Zhang, 2002): la tasa de solapamiento (o proporción de ítems que dos examinados tomados al azar comparten (Way, 1998) y la distribución de la tasa de exposición de los ítems. En principio los algoritmos adaptativos deben favorecer tasas de solapamiento pequeñas y una distribución homogénea de las tasas de exposición de los ítems.

Los primeros métodos propuestos tienen como objetivo fundamental el control de la tasa de exposición máxima de los ítems ($r^{\max}$). Tres de los más importantes son:

- *Método Sympson-Hetter* (Hetter y Sympson, 1997). En este método de control de la exposición se incorpora un parámetro de control para cada ítem [$P(A_j|S_j)$] que determina la probabilidad de aplicación del ítem una vez ha sido seleccionado. Estos parámetros se determinan a través de sucesivos ciclos de simulación de modo que se satisfaga la ecuación:

$$\max[P(A_j)] \le r^{\max} \tag{2.14}$$

Puesto que no es posible aplicar un ítem que no ha sido seleccionado previamente, se cumple que:

$$P(A_j) = P(A_j|S_j)P(S_j) \tag{2.15}$$

Los parámetros $P(A_j|S_j)$ del ciclo $t + 1$ se obtienen de forma que se cumpla: $P(A_j|S_j)_{t+1} P(S_j)_t = P(A_j)_t \le r^{\max}$ (2.16)

Así, el parámetro $P(A_j|S_j)_{t+1}$ toma un valor de 1 cuando la probabilidad de selección del ítem no excede la tasa de exposición máxima y un valor igual a $r^{\max} / P(S_j)_t$ en caso contrario:

$$P(A_j|S_j)_{t+1} = \begin{cases} 1 & si \quad P(S_j)_t \le r^{\max} \\ r^{\max} / P(S_j)_t & si \quad P(S_j)_t > r^{\max} \end{cases} \tag{2.17}$$

Una vez se han obtenido los parámetros $P(A_j|S_j)_{t+1}$ definitivos, estos se incorporan al algoritmo adaptativo. Cuando un ítem es seleccionado, se genera un número aleatorio extraído de una distribución uniforme (0,1), y sólo si ese número es inferior a $P(A_j|S_j)_{t+1}$ el ítem se aplica. En caso contrario, ese ítem no es seleccionado para ese sujeto y se sigue con el siguiente ítem que cumpla el criterio de selección repitiendo el procedimiento. Tres son los problemas principales del método Sympson-Hetter (van der Linden, 2003): (a) es necesario realizar simulaciones previas para calcular los parámetros de control de la exposición; (b) estos parámetros son calculados asumiendo una distribución para el nivel de rasgo; y (c) algunos ítems presentan tasas máximas ligeramente por encima de $r^{\max}$.

- *Método restringido* (Revuelta y Ponsoda, 1998). A diferencia del método anterior, este método no requiere simulaciones previas, sino que los parámetros de control se van ajustando después de cada aplicación. Si la tasa de exposición de un ítem supera la tasa máxima, este ítem deja de estar operativo. La probabilidad de que un ítem *j* sea elegible para el evaluado *N* + 1 se calcula como:

$$P(E_j)_{N+1} = \begin{cases} 1 & si \quad P(A_j)_N < r^{\max} \\ 0 & si \quad P(A_j)_N \geq r^{\max} \end{cases} \tag{2.18}$$

donde $P(A_j)_N$ indica la probabilidad de aplicación en la muestra formada por los *N* sujetos previos.

- *El método de elegibilidad del ítem* (van der Linden y Veldkamp, 2004). Este método tiene la misma lógica que el método anteriormente descrito. En este caso, se parte de que la probabilidad de que un ítem se aplique, $P(A_j)$, depende de la probabilidad de que sea elegible, $P(E_j)$, y de la probabilidad de que se aplique dado que es elegible, $P(A_j \mid E_j)$:

$$P(A_j) = P(A_j \mid E_j)P(E_j) \tag{2.19}$$

Se calcula $P(E_j)_{N+1}$ de forma que pueda mantenerse:

$$P\,(E_j)_{N+1}P(A_j \mid E_j)_N = P(A_j)_N \leq r^{\max} \tag{2.20}$$

Para ello:

$$P(E_j)_{N+1} = \begin{cases} 1 & si \quad P(A_j \mid E_j)_N \leq r^{\max} \\ r^{\max} \big/ P(A_j \mid E_j)_N & si \quad P(A_j \mid E_j)_N > r^{\max} \end{cases} \tag{2.21}$$

Las dos últimas propuestas tienen la ventaja de que no requieren simulaciones para calcular los parámetros de control; además, la selección es independiente de los supuestos sobre la distribución de la habilidad y las tasas de exposición rara vez son mayores a las máximas definidas (Barrada, Abad y Veldkamp, 2009).

Sin embargo, la aplicación de cualquiera de los tres métodos ha revelado que: 1) los ítems con mejores propiedades métricas siguen aplicándose al inicio del test (sólo que en menor proporción); y 2) la forma en la que controla la exposición permite que en niveles de habilidad donde hay pocos sujetos (los niveles extremos) exista un alto grado de solapamiento y aún así las tasas de exposición estén dentro del rango permitido. Para resolver estas limitaciones surgió una segunda generación de métodos de control de la exposición que podrían dividirse en métodos de control de la exposición condicionados a la posición del ítem o condicionados al nivel de habilidad.

Los primeros parten del hecho de que una selección que depende del azar al inicio del test deteriora muy poco la estimación final de los niveles de habilidad (Li y Schafer, 2005). Barrada, Veldkamp y Olea (2009) propusieron el método de múltiples tasas máximas (MRM) por el que: (a) se determinan tantos valores de $r^{\max}$ como ítems se vayan a aplicar en el TAI; (b) los valores de $r^{\max}$ son crecientes a lo largo del test; y (c) al comienzo del test el valor de $r^{\max}$ es igual al mínimo valor admisible.

Por otro lado, la idea fundamental de los métodos condicionados al nivel de habilidad es crear tantos parámetros de control por ítem como intervalos en los que se divide el continuo de habilidad. Se han hecho adaptaciones para el método Sympson-Hetter (Stocking y Lewis, 1998, 2000) y para el método de la elegibilidad del ítem (van der Linden y Veldkamp, 2007).

Los procedimientos antes expuestos, tienen como objetivo aumentar la seguridad del banco disminuyendo la tasa de exposición máxima de los ítems sobre-expuestos, sin cambiar en el proceso el criterio de selección de ítems (usualmente máxima información). Con ello se logra incrementar ligeramente la tasa de exposición de los ítems con propiedades métricas ligeramente inferiores a las de los ítems sobreexpuestos,

pero una gran proporción de ítems del banco continúa presentando bajas tasas de exposición (Barrada, en prensa).

2.5.4. Procedimientos que permiten aumentar la exposición de los ítems infrautilizados

Cuando se utilizan únicamente los criterios de eficiencia en la estimación para la selección de ítems, un alto porcentaje de ítems no se selecciona nunca o prácticamente no se seleccionan. En algunos TAIs este porcentaje llega a alcanzar hasta un 80% del banco (Hornke, 2000). Este hecho no favorece a la seguridad e incrementa considerablemente el coste económico del test.

Para mejorar la gestión del banco de ítems se han propuesto métodos que permiten aumentar el uso de los ítems infraexpuestos. La idea general de estos procedimientos es incrementar la exposición de aquellos ítems con menor capacidad discriminativa aplicándolos al inicio del test, cuando la estimación del nivel de habilidad es imprecisa, e ir incorporando ítems más discriminativos a medida que el test avanza. Algunas de dichas propuestas son:

- *Selección por la distancia entre $\hat{\theta}_{iq}$ y la dificultad del ítem* (Hulin, Drasgow y Parsons, 1983). En éste procedimiento se elimina del criterio de selección el parámetro de discriminación del ítem y se atiende únicamente al parámetro de dificultad. El ítem seleccionado es aquel que minimiza la distancia entre el nivel de rasgo estimado y el parámetro de dificultad del ítem:

$$h_{i(q+1)} = \arg\min_{j \in B_{i(q+1)}} \left| \hat{\theta}_{iq} - b_j \right| \tag{2.22}$$

Las principales ventajas de este método son: 1) que favorece la homogeneidad en las tasas de exposición de los ítems y 2) que selecciona a los ítems de una dificultad adecuada para el nivel de habilidad estimado en ese momento. Sin embargo, es posible que se observe una reducción de la precisión porque no se toma en cuenta la capacidad discriminativa del ítem. Para resolver este problema (Chang y Ying, 1999) propusieron el método alfa-estratificado en el que se combina este criterio con el método estratificado que se describe a continuación.

- *Los métodos estratificados* (Chang y Ying, 1999). En estos métodos el banco presentable ($B_{i(q+1)}$) varía de acuerdo al momento en el que se va a seleccionar el ítem. De forma que al principio $B_{i(q+1)}$ está compuesto por ítems con niveles de discriminación pobres, mientras que al final está compuesto por los ítems con mayor capacidad discriminativa y, por lo tanto, más informativos. El criterio de selección puede ser máxima información o, más usualmente, la selección por la distancia entre el nivel de rasgo estimado y el parámetro *b*. Este procedimiento tiene dos ventajas principales: 1) que se fuerza la selección de los ítems menos discriminativos y 2) se reservan los ítems con mayor discriminación para las fases finales del test.

- *Método progresivo* (Revuelta y Ponsoda, 1998). La principal aportación de éste método es la incorporación de un componente aleatorio en la selección del ítem que favorece la presentación de ítems poco informativos al inicio del test. La función de selección en el método progresivo se compone de dos elementos, uno aleatorio R_j que se calcula en el rango $[0, \max_{j \in B_{i(q+1)}} I_j(\hat{\theta}_{iq})]$ y un segundo elemento que es la información de Fisher $[I(\hat{\theta}_q)]$. Ambos elementos están ponderados de manera que al principio del test la selección depende principalmente del componente aleatorio, mientras que a medida que la prueba avanza se va incrementando el peso de la información en la selección del ítem. Concretamente, el ítem seleccionado es el que hace máxima la siguiente función:

$$h_{i(q+1)} = \arg\max_{i \in B_{i(q+1)}} [(1 - W_{q+1}) R_j + W_{q+1} I_j(\hat{\theta}_{iq})] \tag{2.23}$$

  donde W_q varía según la posición del ítem ($q+1$) en el test y se define como:

$$W_{q+1} = \frac{q}{Q} \tag{2.24}$$

  donde *Q* es la longitud prefijada del TAI.

  Revuelta y Ponsoda (1998) demostraron que este método aumenta la exposición de los ítems infrautilizados con poca o ninguna pérdida en la precisión.

- *Método progresivo con parámetro de aceleración* (Barrada, Olea, Ponsoda, Abad, 2008). Una posible modificación del método progresivo es flexibilizar el componente aleatorio utilizando una función de cambio que no es necesariamente lineal y que se calcula como:

$$W_{q+1} = \begin{cases} 0 & si \quad q+1=1 \\ \dfrac{\sum_{z=1}^{q+1}(z-1)^t}{\sum_{z=1}^{Q}(z-1)^t} & si \quad q+1 \neq 1 \end{cases} \tag{2.25}$$

donde el parámetro *t* permite establecer la velocidad con la que se reduce el componente aleatorio.

- *Método proporcional con parámetro de aceleración* (Barrada, et al., 2008): La propuesta de Barrada et al. (2008) es una modificación del método proporcional propuesto por Segall (2004). En este procedimiento, la función de información del ítem sirve para determinar la probabilidad de seleccionarlo, $P(S_j)$:

$$P(S_j) = \frac{V_j}{\sum_{j'=1}^{J} V_{j'}} \tag{2.26}$$

donde V_j es la función de valoración. Por ejemplo:

$$V_j = \left(1 - m_{ij}\right) I_j(\hat{\theta}_{iq}) \tag{2.27}$$

Una vez que se ha determinado la probabilidad para todos los ítems se construye una distribución acumulativa de probabilidades. Un número aleatorio extraído extraído dentro del intervalo uniforme (0, 1) determina qué ítem es seleccionado. Este procedimiento reduce considerablemente la varianza de la tasa de exposición. El problema es que la probabilidad de selección de un ítem *j* se mantiene constante a través del test. Sin embargo, como ya se ha mencionado, mientras que es ventajoso favorecer la selección de ítems cuyo parámetro de discriminación sea bajo, al inicio del test, esto debe cambiar a medida que el test trascurra. Con este objetivo, Barrada et al. (2008) proponen que a la probabilidad de selección venga determinada por la información de Fisher elevada a una potencia de valor creciente según avanza el test, lo cual permite que el componente de información vaya adquiriendo preponderancia a medida que el test progresa. Para ello se calcula:

$$V_{j(q+1)} = \left(1 - m_{ij}\right) I_j(\hat{\theta})^{P_{q+1}} \tag{2.28}$$

Y el exponente de la función P_{q+1} se define mediante la siguiente fórmula:

$$P_{q+1} = \begin{cases} 0 & si \quad q+1=1 \\ \dfrac{Q\sum_{z=1}^{q+1}(z-1)^t}{\sum_{z=1}^{Q}(z-1)^t} & si \quad q+1 \neq 1 \end{cases} \tag{2.29}$$

Los resultados de la investigación muestran que para evitar el problema de la infraexposición resulta recomendable la aplicación de alguno de los métodos propuestos en este apartado. En relación a los métodos con parámetro de aceleración, se encuentra que la elección de este parámetro viene condicionada por los objetivos del evaluador. El decremento en la precisión, el incremento en el control de la exposición y la mejora del mantenimiento del banco vienen determinados por la magnitud del parámetro de aceleración. Si lo importante es la precisión se deben utilizar valores de *t* bajos. Por el contrario valores de *t* altos permiten mantener durante una parte importante del test un alto componente de azar en la selección, lo que favorece a la seguridad y al mantenimiento del test.

2.6. La evaluación de la fiabilidad y la validez en los TAIs

El desarrollo tecnológico implicado en los tests informatizados ha transformado en muchos aspectos la medición. Así, en la actualidad, los test informatizados hacen posible la evaluación de rasgos que sería imposible evaluar mediante tests de lápiz y papel. Algunos ejemplos de esto son pruebas de selección de controladores aéreos que permiten medir habilidades complejas de coordinación viso-motora (p.e., MICROPAT), pruebas de certificación para arquitectos en las que se debe desarrollar un diseño por ordenador bajo algunas restricciones, pruebas de inglés para evaluar el nivel hablado en las que se emiten las respuestas a través de un micrófono (p.e., TOEFL) o test de evaluación educativa en las que es posible escribir números, palabras, ecuaciones, elegir en pantallas táctiles, dibujar, etc.

No obstante no debemos olvidar que, más allá de la validez aparente que suele suponer el uso de estas nuevas tecnologías, se deben exigir para las puntuaciones obtenidas en Tests informatizados, las mismas garantías psicométricas, de fiabilidad y validez, que en tests de lápiz y papel. Así, antes de que un TAI se encuentre operativo, pero también durante su vida útil, es común que se realicen controles psicométricos de

calidad en los que se evalúa la precisión en la medida y se estudia la validez de las puntuaciones, acumulando evidencia para apoyar las inferencias que se hacen sobre las puntuaciones del TAI (Olea y Ponsoda, 2003). Se han ofrecido algunas directrices específicas aplicables al desarrollo de pruebas de evaluación informatizadas (ITC, 2000).

En relación a la fiabilidad, la expansión de los modelos psicométricos de TRI ha dotado a los psicómetras de un marco adecuado para evaluar los niveles de precisión. Sin embargo, la existencia de estos modelos no resuelve todos los problemas. Puesto que un TAI requiere la construcción de un banco amplio de ítems, se hace necesario utilizar un diseño de anclaje para calibrar los ítems. Una vez estimados los parámetros de los ítems y contrastados los supuestos del modelo (p.e., unidimensionalidad), se deben llevar a cabo estudios de simulación que sirven para describir la precisión del TAI en cada nivel de habilidad, considerando distintas alternativas el algoritmo adaptativo (p.e., variando el criterio de parada, el método o los parámetros de control de la exposición, etc.).

En relación a la validez, los TAIs plantean nuevas situaciones como la aplicación informatizada o la obtención de respuestas de forma adaptativa, que generan nuevas preguntas y desafíos en la evaluación de la validez. Parte de esa especificidad tiene que ver con la aplicación informatizada. Por ejemplo, debe garantizarse que las condiciones de aplicación están estandarizadas, lo que puede depender de unos requisitos mínimos en relación a las características del ordenador y/o, en las aplicaciones vía web, de la conexión. También debe controlarse que destrezas específicas relacionadas con el uso de los ordenadores no afecten a la ejecución de la tarea.

La obtención de respuestas de forma adaptativa plantea otras cuestiones adicionales como, por ejemplo, si debe permitirse la revisión de las respuestas, cómo controlar el balance de contenidos en el TAI o cuál es el efecto en la motivación de ajustar el nivel de dificultad al nivel de habilidad del evaluado.

Finalmente, más allá de las potenciales amenazas a la validez propias de la aplicación adaptativa informatizada, es necesario reconocer que el proceso de estudio de las propiedades psicométricas de los ítems se complica en un TAI ya que: 1) puesto que el TAI se suele utilizar en procesos de evaluación continua, los ítems también deben ser sometidos a controles psicométricos continuos y, adicionalmente, se requiere la

actualización de los ítems; 2) en las bases de datos para realizar los análisis casi siempre hay valores perdidos.

Resulta útil distinguir tres momentos en el estudio de las propiedades psicométricas de un ítem en el TAI:

- *Fase de construcción del banco:* Se estudian por primera vez las propiedades psicométricas de los ítems que compondrán el banco. El análisis se realiza en un contexto en el que ningún ítem se aplica adaptativamente pero, usualmente, no todos los evaluados responden a todos los ítems (i.e., se aplica un diseño de anclaje). Los datos perdidos son MCAR (*missing completely at random*). No es infrecuente que en esta fase los ítems se apliquen en un formato de lápiz y papel.
- *Fase de mantenimiento:* Se comprueban las propiedades psicométricas de los ítems operativos en la aplicación adaptativa. Las propiedades métricas del ítem pueden deteriorarse debido sobre todo a la sobreexposición, a cambios en el formato de aplicación (informatizado vs lápiz y papel) o porque se vuelvan obsoletos. El análisis se realiza en un contexto en el que todos los ítems se aplican adaptativamente, por lo que los valores perdidos no son MCAR, existiendo restricción de rango en el nivel de rasgo de los evaluados que responden a cada ítem.
- *Fase de actualización del banco:* El banco de ítems debe actualizarse de forma que se sustituyan los ítems que se van analizando. Una situación usual es que se deterioren primero las propiedades psicométricas de los ítems más discriminativos, que, dadas las características del algoritmo adaptativo, serán más expuestos. Esto supone que en el proceso de actualización se demanda una alta capacidad para construir ítems igual de buenos que los ítems que se van eliminando. En relación al análisis, los nuevos ítems que se añaden al banco (pretest) se aplican junto al TAI operativo. Por tanto, el nuevo ítem no se aplica adaptativamente pero sus propiedades psicométricas deben estudiarse en un contexto en el que el resto de los ítems se han aplicado adaptativamente.

En el capítulo siguiente se abordan de forma específica los problemas que supone cada una de estas situaciones en el estudio del funcionamiento diferencial de los ítems.

Capítulo 3

El estudio del DIF en Test Adaptativos Informatizados

3.1. Introducción

Embretson (2004) señala que en la segunda era de los test psicométricos muchos de los procedimientos clásicos de validez, incluyendo los estudios de DIF, deben adaptarse a las nuevas formas de construcción de test. Entre estas nuevas formas están los tests adaptativos informatizados (TAIs).

La detección del DIF es más importante en un TAI que en un test convencional (Zwick, 2000, 2007). En primer lugar, la longitud de un test adaptativo suele ser en torno a la mitad de la longitud del test convencional al que substituye. Por lo tanto, la presencia de un ítem con DIF puede afectar en mayor medida a la estimación del rasgo. En segundo lugar, la aplicación de un ítem con DIF puede rebajar la idoneidad de los siguientes ítems administrados en el TAI, ya que la selección de ítems depende las respuestas dadas por el evaluado y los parámetros estimados en los ítems previamente administrados. Finalmente, la aplicación de un test en un medio informatizado puede crear fuentes potenciales de DIF. Por ejemplo, las diferencias que ocurren entre algunas poblaciones en relación a la familiaridad en el uso de los ordenadores (Powers y O´Neill, 1993) o a la ansiedad ante la imposibilidad de poder revisar las respuestas (Olea, Revuelta, Ximénez y Abad, 2000).

La evaluación adaptativa introduce algunas dificultades particulares para el análisis del DIF (Lei, Chen y Yu, 2006; Nandakumar y Roussos, 2001, 2004; Steinberg, Thissen y Wainer, 1990; Zwick, 2000, 2007):

1) En los métodos OCI, pierde sentido usar la suma de respuestas correctas como variable de igualación. De hecho, en un TAI se espera que los evaluados acierten aproximadamente la misma proporción de ítems con independencia de su nivel de habilidad.
2) Existe una relación entre la dificultad de los ítems y el nivel de habilidad de las personas que responden a un ítem que se aplica de forma adaptativa. Por lo tanto, existe restricción de rango en la distribución del rasgo de los evaluados que responden a un ítem. Además, esta distribución varía de ítem a ítem.
3) Las matrices de datos obtenidas son incompletas, debido a que diferentes sujetos responden a ítems distintos. Cada grupo recibirá ítems ajustados en dificultad a la distribución del nivel de rasgo. Esto puede hacer que haya poco solapamiento en los ítems de anclaje o en los niveles de la variable de igualación.
4) A pesar de que el TAI se aplique a una muestra considerable, el número de personas que responde a algunos ítems puede ser pequeño, dificultando la estimación de sus parámetros.
5) El efecto de la contaminación en un TAI puede ser mayor que en un test fijo, ya que suelen ser más cortos.
6) Aunque de relevancia menor, un problema práctico puede ser el tamaño de las matrices de respuesta. En un test fijo raramente se analizan más de 50 ítems simultáneamente. En un TAI, pueden llegar a analizarse bancos con más de 1000 ítems. Por tanto, la eficiencia computacional es un criterio muy relevante para elegir entre los distintos métodos de detección del DIF.

La presencia de estas dificultades depende de en qué momento se evalúa el DIF de un ítem a lo largo de la vida de un TAI. Pueden distinguirse tres momentos:

1) Creación del banco de ítems. En esta fase se calibran por primera vez los ítems que conformarán el banco del TAI. Para ello, la aplicación de los ítems no es adaptativa. Usualmente se utiliza algún diseño de anclaje. En este contexto es posible aplicar las técnicas tradicionales de análisis del DIF.
2) Actualización del banco de ítems. En esta fase se calibran ítems nuevos (pretest) para aumentar o mantener el tamaño del banco. Los ítems nuevos están en proceso de prueba (ítems pretest) y son analizados de manera exhaustiva antes de convertirse en ítems operativos, ya que eliminarlos en esta fase tiene menos

consecuencias que hacerlo en una fase operativa (Lei et al., 2006; Nandakumar y Roussos, 2001, 2004). Los ítems pretest se aplican a todos los evaluados, durante la aplicación del TAI, y no se emplean para puntuar a los evaluados. En esta situación se genera una matriz dividida en dos secciones: (a) las respuestas a los ítems operativos que conforman una matriz de datos incompleta; y (b) las respuestas a los ítems pretest, de los que se pretende estudiar el DIF. En esta segunda sección están las respuestas de todos los evaluados, por lo que no existe restricción del rango de habilidad para estos ítems, aunque sí se presentan el resto de las dificultades mencionadas.

3) Revisión del banco de ítems. En este caso se revisan los parámetros de los ítems operativos. La dificultad estriba en que el ítem del que se estudia el DIF se aplica de forma adaptativa. En esta situación, se tiene para el análisis del DIF una matriz de datos incompleta, restricción de rango en el nivel de habilidad para el análisis de cada ítem y, adicionalmente, si varios ítems aplicados en el TAI tienen DIF, una estimación errónea del nivel de habilidad.

Hasta el momento cinco de los métodos de detección de DIF en tests fijos han sido adaptados a la evaluación del DIF en TAIs. Zwick, Thayer y Wingersky (1993, 1994a, 1994b, 1995) realizaron una adaptación del estadístico de Mantel-Haenzsel ($\hat{\Delta}_{MH}$) y del procedimiento de Estandarización (STD). En trabajos posteriores propusieron una aproximación bayesiana a dichos estadísticos (Zwick, Thayer y Lewis, 1997, 1999; Zwick y Thayer, 2002, 2003). En estos trabajos se estudian básicamente la detección del DIF en ítems operativos. Por otro lado, Nandakumar y Roussos (2001, 2004) propusieron el CATSIB, que es la versión adaptada del SIBTEST. Lei et al. (2006) adecuaron la Regresión Logística (RL) y el test de razón de verosimilitudes de la TRI (IRT LRT). Tanto Nandakumar y Roussos como Lei et al. se han centrado en el estudio de DIF en ítems pretest. En los siguientes apartados se describen sucintamente algunas de las modificaciones necesarias para adaptar estos métodos.

3.2. Métodos adaptados para la evaluación de DIF en TAIs

3.2.1. Métodos basados en la TRI: IRT LR test

Steinberg et al. (1990) consideran que el test de razón de verosimilitudes para la comparación de modelos (IRT LRT) podría implementarse fácilmente para el análisis del DIF en TAIs, puesto que la TRI forma parte del proceso de construcción de un TAI. Como se describió en el capítulo 1, el IRT LRT permite contrastar si las CCI difieren significativamente entre los grupos y es uno de los procedimientos eficaces en la identificación del DIF tanto unidireccional como no unidireccional. Algunas ventajas conceptuales de esta técnica son:

1) La necesidad de muestras grandes y la comprobación del ajuste al modelo, que suelen considerarse desventajas de la TRI, son requisitos necesarios en el desarrollo de un TAI. La muestra para evaluar las propiedades psicométricas de un ítem pretest siempre deben ser suficientemente grandea como para calibrar el ítem eficazmente. Además, si las respuestas al ítem no se ajustan al modelo el ítem deberá ser descartado. En condiciones adecuadas, y si se cumplen los supuestos del modelo de TRI, esta estrategia puede resultar óptima por su mayor potencia si la muestra es grande (French y Finch, 2008; Lopez-Rivas, Stark, y Chernyshenko, 2008; Stark, Chernyshenko, y Drasgow, 2006; Wang y Yeh, 2003; Woods, 2008a).
2) En función de los contrastes específicos sobre los parámetros *a, b* y *c*, el uso de esta estrategia permite obtener información precisa del tipo de DIF.
3) Se dispone de *software* que permite la aplicación sencilla de esta técnica, especialmente para el contexto de tests fijos. El programa IRTLRTDIF v.2.0b (Thissen, 2001) automatiza las múltiples comparaciones de modelos anidados que se llevan a cabo cuando se sigue esta estrategia.

Por otro lado, la aplicación de IRT LRT no está exenta de problemas. La enorme proporción de respuestas faltantes, el escaso rango en el nivel de habilidad de las personas que responden al ítem y, si los grupos difieren en habilidad, la baja proporción de ítems comunes aplicados a los dos grupos puede dificultar obtener los parámetros de

los ítems del TAI operativo. Adicionalmente, en el contexto de los TAIs se requiere la estimación de los parámetros de un banco grande de ítems y el problema del coste computacional puede ser relevante, especialmente si se considera la posibilidad de añadir una fase de depuración.

Recientemente, Lei et al. (2006) propusieron substituir la matriz observada de datos (matriz completa para los ítems pretests y dispersa para los ítems operativos) por una matriz imputada (completa para todos los ítems). Para ello, las respuestas a los ítems no administrados eran imputadas a partir del nivel de rasgo estimado en el TAI y los parámetros de los ítems. Sin embargo, aunque se resuelve el problema de convergencia en la calibración de ítems, su propuesta resulta algo ineficiente por el alto coste computacional (i.e., se requiere la estimación de los parámetros en los ítems operativos) y por las limitaciones del IRTLRDIF (Thissen, 2001), que limita el número de ítems que se puede analizar (p.ej., el número de ítems operativos para el anclaje está limitado a 200).

3.2.2. Métodos OCI aplicados a la detección del DIF en TAIs

El primer problema a resolver por los métodos OCI es el uso de una variable de igualación distinta a la suma de aciertos, puesto que en un TAI ésta no es una estimación adecuada del nivel del rasgo. Así, las primeras investigaciones se dirigieron a la búsqueda de una variable de igualación válida (Zwick et al., 1993, 1994a). Zwick et al. (1993, 1994a) siguen la sugerencia de Steinberg et al. (1990) de usar como variable de igualación la puntuación esperada en el banco de ítems:

$$W_0 = E(X \mid \hat{\theta}_{TAI}) = \sum_{j=1}^{J} P_j(u_j = 1 \mid \hat{\theta}_{TAI}) \tag{3.1}$$

donde $P_j(u_j = 1 \mid \hat{\theta}_{TAI})$ es la función de respuesta estimada para el ítem j, asumiendo los parámetros de los ítems estimados y evaluada en $\hat{\theta}_{TAI}$, $\hat{\theta}_{TAI}$ es la estimación de la habilidad basada en la respuesta a los ítems aplicados adaptativamente y J es el número de ítems en el banco. Comúnmente la puntuación esperada se redondea a un valor entero para formar los intervalos de igual tamaño.

Una cuestión clave es cómo incluir el ítem l analizado en el cálculo de la variable de igualación. En contra de lo que se podría pensar, incluir el ítem para el cálculo puede resultar necesario para aquellos procedimientos que utilizan la puntuación

directa como variable de igualación (Monahan y Ankenmann, 2010; Walker, Beretvas y Ackerman, 2001). Con tests fijos la no inclusión del ítem analizado en la puntuación directa infla las tasas de error Tipo I si los test son cortos, poco fiables o si los grupos difieren en la distribución del nivel de rasgo. Para corregir este problema, algunos procedimientos de DIF en test fijos (MH, STD, RL) añaden la puntuación en el ítem que se está estudiando a la puntuación directa (Holland y Thayer, 1988). En el contexto de los TAIs, Zwick et al. (1993, 1994b) consideran varias formas de añadir el ítem analizado en la variable de igualación:

1) Si se analiza el DIF de un ítem operativo, puede utilizarse W_0, ya que el ítem estudiado es un ítem del TAI. En el análisis del DIF de un ítem pretest, puede utilizarse cualquiera de las siguientes estrategias.
2) W_1, que se calcula como la puntuación verdadera esperada en el banco de ítems más la puntuación en el ítem pretest estudiado (X_l):

$$W_1 = \sum_{j=1}^{J} P_j(u_j = 1 \mid \hat{\theta}_{TAI}) + X_l \tag{3.2}$$

3) W_2, o "variable de igualación teóricamente óptima", que se calcula como la puntuación esperada en el banco de ítems a partir de $\hat{\theta}_{TAI+l}$, donde $\hat{\theta}_{TAI+l}$ es la estimación ML basada en el TAI y en el ítem pretest (l), más la puntuación verdadera estimada en el ítem pretest, $P_l(u_l = 1 \mid \hat{\theta}_{TAI+l})$:

$$W_2 = \sum_{j=1}^{J} P_j(u_j = 1 \mid \hat{\theta}_{TAI+l}) + P_l(u_l = 1 \mid \hat{\theta}_{TAI+l}) \tag{3.3}$$

4) W_3, que es una aproximación a la "variable de igualación teóricamente óptima". Esta aproximación es similar a la anterior, con la diferencia de que la estimación de la habilidad sólo se basa en las respuestas al TAI ($\hat{\theta}_{TAI}$).

$$W_3 = \sum_{j=1}^{J} P_j(u_j = 1 \mid \hat{\theta}_{TAI}) + P_l(u_l = 1 \mid \hat{\theta}_{TAI}) \tag{3.4}$$

W_1 sería la aproximación más adecuada cuando se analiza el DIF de un ítem pretest del que se desconocen los parámetros. W_2 y W_3 son aproximaciones más complejas porque requieren la estimación de los parámetros en el ítem pretest l. Por tanto, pueden considerarse más adecuadas para la estimación del DIF en ítems pretest

para los que se tiene una estimación provisional de sus parámetros. Una vez calculada la variable de igualación pueden obtenerse medidas de DIF como $\hat{\Delta}_{MH}$ o el STD (Zwick et al., 1993, 1994a).

Otra alternativa es trabajar directamente con $\hat{\theta}_{TAI}$ como variable de igualación. Esta estrategia se ha propuesto para aplicar la regresión logística para la detección del DIF en TAIs en ítems pretest (Lei et al., 2006).

Finalmente, Nandakumar y Roussos (2001, 2004) consideran que $\hat{\theta}_{TAI}$ puede ser un estimador sesgado de θ y proponen una corrección análoga a la corrección por regresión propuesta por Shealy y Stout (1993b) en el programa SIBTEST. En el programa CATSIB, que es la adaptación de SIBTEST para la detección de DIF en TAIs, la variable de igualación corregida, $\hat{\theta}^{*}_{TAI}$, se calcula como:

$$\hat{\theta}^{*}_{TAI} = E(\theta \mid \hat{\theta}_{TAI}) = \mathrm{E}(\hat{\theta}_{TAI}) + \rho^{2}(\hat{\theta}_{TAI} - \mathrm{E}(\hat{\theta}_{TAI})) \tag{3.5}$$

donde ρ representa la correlación entre $\hat{\theta}_{TAI}$ y θ. ρ se obtiene para cada grupo como:

$$\rho = \sqrt{1 - \frac{s_e^2}{s^2_{\hat{\theta}_{TAI}}}} \tag{3.6}$$

donde s_e^2 y $s^2_{\hat{\theta}_{TAI}}$ son, respectivamente, el promedio de la varianza del error de estimación de θ y la varianza de $\hat{\theta}_{TAI}$. La puntuación de cada evaluado se calcula teniendo en cuenta el grupo *g* al que pertenece:

$$\hat{\theta}^{*}_{TAI(g)} = \mathrm{E}_g(\theta \mid \hat{\theta}_{TAI}) = \mathrm{E}_g(\hat{\theta}_{TAI}) + \rho_g^2\left[\hat{\theta}_{TAI} - \mathrm{E}_g(\hat{\theta}_{TAI})\right] \tag{3.7}$$

Esta corrección es equivalente a la propuesta por Kelley (1923, citado en Monahan, y Ankenman, 2010) en el contexto de la teoría clásica de tests para corregir las puntuaciones observadas considerando el error de medida. En el CATSIB los evaluados se agrupan según la puntuación corregida, $\hat{\theta}^{*}_{TAI}$. La fórmula para el estimador de β en CATSIB si analizamos el ítem *l*, es:

$$\hat{\beta}_{l(CATSIB)} = \int (P_{lR}(u_l = 1 \mid \hat{\theta}^{*}_{TAI}, g = R) - P_{lF}(u_l = 1 \mid \hat{\theta}^{*}_{TAI}, g = F)) f(\hat{\theta}^{*}_{TAI}) d(\theta) \tag{3.8}$$

donde $P_{lR}(u_l = 1 \mid \hat{\theta}^{*}_{TAI}, g = R)$ es la proporción observada de aciertos para los evaluados del grupo de referencia que tienen una habilidad estimada $\hat{\theta}^{*}_{TAI}$ y $f(\hat{\theta}^{*}_{TAI})$ es la proporción de evaluados que se encuentran en ese nivel. Debido a que $\hat{\theta}^{*}_{TAI}$ es un valor

real, difícilmente los evaluados tienen el mismo valor en esa variable. Por lo tanto, el rango observado de $\hat{\theta}^{*}_{TAI}$ se divide en *K* intervalos del mismo tamaño. Una vez calculada la variable de igualación, los procedimientos mencionados (Mantel-Haentszel, Regresión logística y CATSIB) resuelven varios de los problemas señalados:

1) Desaparece el problema de los valores perdidos (todos los evaluados que han respondido al ítem *l* tendrán una puntuación en la variable de igualación).
2) La restricción del rango de habilidad para los ítems operativos de anclaje ya no es un problema, puesto que esa información sólo se utiliza para calcular la variable de igualación.
3) También se atenúa el potencial problema de que tiendan a aplicarse distintos ítems de anclaje en grupos distintos.

Todavía puede seguir resultando un problema: (1) la restricción de rango en el nivel de rasgo cuando el ítem estudiado se aplica adaptativamente; (2) la presencia de contaminación en el TAI, ya que incluir un paso de depuración supondría un alto coste computacional si se requiere el análisis de todos los ítems del banco, considerando muestras y variables de igualación distintas para cada ítem.

Otro problema que no está bien resuelto es el modo de establecer los intervalos de rasgo en MH, CATSIB o STD (Rousssos, Nandakumar y Banks, 2006). Si se escoge un número demasiado grande, habrá un problema de dispersión de los datos, puesto que habrá celdas en la tabla de contingencia con pocos evaluados que serán eliminadas de los análisis (p.e., en CATSIB) y se reducirá la potencia. Si se escoge un número demasiado pequeño, las tasas de error Tipo I se inflarán. En el caso extremo de un único intervalo, la presencia de DIF se confunde con la presencia de impacto.

El número óptimo de intervalos puede depender de si se analiza un ítem pretest o un ítem operativo. En el primer caso, los evaluados se distribuyen en un rango amplio de niveles de rasgo mientras que el segundo se sitúan en un rango estrecho de habilidad. Por ejemplo, Zwick y Thayer (2002, 2003) proponen, en el contexto de detección de ítems operativos, el uso de intervalos de amplitud 2 cuando se analizan la variable de igualación W_0, ya que obtenían mejores resultados que cuando se utilizaban intervalos de amplitud 1. Sin embargo, aplicado ese criterio en el análisis de ítems pretest, llevará

a la presencia de celdas en la tabla de contingencia con frecuencias escasas o nulas, debido a la dispersión de los datos.

En el contexto de la detección de ítems pretest, Nandakumar y Roussos (2001) proponen para CATSIB comenzar con un número alto de intervalos (80) y evaluar el porcentaje de evaluados perdidos cuando se eliminan los intervalos con menos de 3 evaluados. Se reduce de uno en uno el número de intervalos hasta que la proporción de evaluados eliminados sea menor que 7.5% en ambos grupos o hasta que se llegue a un mínimo de 20 intervalos. Roussos et al. (2006) sugieren que el uso de un mínimo de 10 intervalos puede ser más apropiado, resultando en una menor perdida de sujetos. Además encuentran que el uso de intervalos basados en los percentiles puede proporcionar mejores resultados que los intervalos basados en intervalos de igual amplitud en el rasgo. Lei et al. (2006) también encuentran mejores resultados con 10 que con 20 intervalos cuando los tamaños muestrales son distintos entre los grupos.

3.2.3. Comparación de métodos

En el Capítulo 1 se describieron algunos de los resultados empíricos obtenidos en tests fijos en relación a los distintos métodos de detección de DIF. Entre los métodos clásicos, MH, SIBTEST y RL resultan apropiados para la detección del DIF unidireccional. Sin embargo, en presencia de impacto, si el test es corto y los ítems varían en discriminación, pueden inflarse las tasas de error Tipo I. En ese caso, SIBTEST es el método que mejor controla las tasas de error Tipo I, gracias a la corrección por regresión (Bolt, 2000; Gierl, Gotzmann y Boughton, 2004; Jiang y Stout, 1998). Esta corrección parece mejorar también el rendimiento de todas las técnicas (DeMars, 2009). Por otro lado, RL es la técnica clásica más adecuada para la detección del DIF no unidireccional (Narayanan y Swaminathan, 1996), siempre y cuando los datos se ajusten al modelo logístico de dos parámetros (Li y Stout, 1996).

En el contexto de los TAIs, Zwick et al. (1993, 1994a, 1994b) estudiaron el uso de $\hat{\Delta}_{MH}$ y SPD para la detección del DIF unidireccional en un TAI tanto para ítems pretest (1994b) como para ítems aplicados de forma adaptativa (1993, 1994b). Las respuestas se generaban mediante un modelo logístico de tres parámetros (3PL). En sus estudios de simulación se distinguen dos fases. En la fase de calibración se simulan las respuestas a los 75 ítems operativos en una muestra de 2000 evaluados. Se simulaban

tres condiciones: (a) banco sin DIF (banco A); (b) banco de ítems con DIF (banco B); (c) banco de ítems con DIF correlacionado con la dificultad (banco C). En la fase de aplicación del TAI, se simula un TAI de 25 ítems utilizando los parámetros estimados en la primera fase. La selección de ítems se realiza por el método de máxima información y la estimación del nivel de rasgo, por máxima verosimilitud. Finalmente, en el estudio de ítems pretest, se simulaban, para esos mismos sujetos, las respuestas a 15 ítems pretest. En todos los casos, la distribución del rasgo en el grupo de referencia seguía la distribución normal estándar.

Los factores manipulados y sus niveles son, en parte, comunes a los distintos estudios: (a) tamaño de las muestras (igual: N_R =500, N_F = 500; distinto: N_R =900, N_F = 100); (b) tamaño del impacto, manipulando la distribución del grupo focal [N(0, 1); N(0.5, 1); N(–1,1)]; y (c) tamaño del DIF que depende del parámetro d_j [–0.70, –0.35, 0, 0.35, 0.70], siendo $d_j = b_{jR} - b_{jF}$.

En los resultados se analizaron las correlaciones entre la magnitud verdadera del DIF (un valor proporcional a $a_j d_j$) y los estadísticos del DIF para $\hat{\Delta}_{MH}$, así como el error estándar de estimación de dichos estadísticos ($SE_{\Delta_{MH}}$ y SE_{STD}) cuando se emplearon, entre otras, las variables de igualación W_0 (para los ítems operativos) y W_1, W_2 y W_3 (para los ítems pretest). Se encontró que el cálculo de la variable de igualación más sencillo (W_1) proporcionaba resultados similares a los procedimientos más sofisticados (W_2 o W_3).

Los autores señalan correlaciones superiores a 0.91 entre la magnitud de DIF verdadera y los estadísticos estimados ($\hat{\Delta}_{MH}$ y SPD) y una buena clasificación de los ítems según el tamaño del DIF (A, B, C), incluso en las condiciones con impacto, especialmente en los bancos A y B.

El error típico del estadístico $\hat{\Delta}_{MH}$ era menor en la detección del DIF en ítems operativos, puesto que el error típico de $\hat{\Delta}_{MH}$ disminuye cuando la proporción de aciertos se aproxima a 0.5. Los autores concluyen que la detección del DIF puede ser más fácil cuando el ítem se aplica adaptativamente.

Zwick et al. (1995) extienden los estudios anteriores, aplicando un modelo de Rasch para estimar los parámetros en una muestra de calibración generada con un modelo de 3PL. Los parámetros estimados mediante el modelo de Rasch se usaron para la selección de los ítems (por máxima información) y la estimación de la habilidad. Los

resultados señalan que el empleo de un modelo incorrecto en la estimación conduce a una disminución en la identificación del DIF. Por otro lado Zwick et al. (1997, 1999) y Zwick y Thayer (2002, 2003) han mostrado que la obtención de la distribución posterior de los estadísticos MH (por métodos bayesianos), puede ser utilizada para mejorar la tasa de detección del DIF y obtener la probabilidad posterior de que el ítem sea clasificado en una categoría de DIF (A, B o C).

A pesar de que los anteriores resultados son prometedores, una limitación general de los estudios realizados con MH es que no se obtienen las tasas de error Tipo I ni de potencia, ni tampoco se compara MH con otras técnicas, por lo que es difícil concluir sobre su funcionamiento en relación a los otros métodos.

Nandakumar y Roussos (2001, 2004) examinaron la efectividad del CATSIB para detectar el DIF unidireccional en ítems pretest. Se simuló un banco de 1000 ítems ajustados al modelo 3PL, con una distribución de parámetros realista similar a la del banco del LSAC (Law School Admission Council). Se simularon las respuestas a un TAI de 25 ítems, seleccionados con un método basado en la propuesta de Kingsbury y Zara (1991); durante los 9 primeros ítems aplicados se seleccionaba el ítem *k* que se encuentra entre los (10–*k*) ítems más informativos del banco para el nivel de rasgo estimado. A partir del ítem 10 se seleccionaba el ítem más informativo. Al finalizar la aplicación adaptativa se administraban los ítems pretest. En todos los casos, la distribución del rasgo en el grupo de referencia seguía la distribución normal estándar.

En su estudio se manipularon los siguientes factores, entre otros: (a) el tamaño de las muestras (500–500, 500–250 y 250–250); (b) tamaño del impacto, manipulando la distribución del grupo focal ($N \sim (0,1)$; $N \sim (-0.5,1)$; $N \sim (-1,1)$); y (c) β o tamaño del DIF (0 –en 6 ítems–, 0.5 –en 5 ítems– y 0.10 –en 5 ítems–).

Los resultados del estudio mostraron un adecuado control del error Tipo I en los casos en los que se empleó la corrección en la estimación del rasgo. Cuando no se realizó la corrección, en el 75% de los casos la tasa de error Tipo I estuvo fuera del intervalo para el nivel nominal aceptado. Respecto a la potencia, se estimó que para muestras grandes (500–500 y 500–250), las tasas de potencia de los ítems eran mayores a 0.80 cuando el DIF fuera grande, $\beta > 0.10$. En muestras pequeñas (250-250), la potencia promedio fue de 0.73 si $\beta > 0.10$ y disminuyó a 0.64 en la condición de mayor impacto. cuando la distribución del grupo focal fue $N \sim (-1,1)$. Las tasas de potencia también variaron dependiendo de las características de los ítems; la potencia fue mayor

para los ítems muy discriminativos. Los autores señalan que el empleo de $\hat{\beta}$ es bastante adecuado como estimador del grado de DIF porque permite hacer juicios sobre si mantener o no un ítem en futuras aplicaciones. La aplicación de CATSIB con regresión mostraba buenos resultados incluso en presencia de impacto.

Lei et al. (2006) comparan la efectividad de tres métodos para detectar el DIF en ítems pretest. Los tres métodos comparados fueron el CATSIB (con corrección por regresión), la versión modificada de la regresión logística en el cual la puntuación directa se sustituye por la estimación de la habilidad basada en la TRI ($\hat{\theta}_{TAI}$), y su propuesta de modificación del IRT LRT, basado en la imputación de respuestas. Se simuló un banco de 360 ítems ajustados al modelo 3PL, simulando las condiciones del ACT Math que evalúa seis áreas de contenido. Se simulaban las respuestas a un TAI de 30 ítems, con selección de ítems por máxima información, incorporando el método de control de exposición y balanceo de contenidos propuesto por Chen, Ankenmann y Chang (2000). El nivel de rasgo se estimaba por el procedimiento bayesiano EAP. Además del TAI se aplicaban 16 ítems pretest, que variaban según la condición: (a) sin DIF; (b) DIF unidireccional; y (c) DIF no unidireccional. El parámetro *c* era constante en todas las condiciones (0.15). Las respuestas en el grupo de referencia seguían una distribución normal estándar.

En el caso de CATSIB y RL, los autores utilizaron $\hat{\theta}^{*}_{TAI}$ y $\hat{\theta}_{TAI}$, respectivamente como variables de igualación. En el caso de IRT LRT, se tomaron como ítems de anclaje un subconjunto de 84 ítems operativos, para permitir el análisis con IRTLRDIF v2.0b (Thissen, 2001).

Se manipularon los siguientes factores: (a) tamaño de las muestras (iguales: N_R = 500; N_F = 500; diferentes: N_R = 900; N_F = 100); (b) presencia de impacto, manipulando la distribución del grupo focal (sin impacto; con impacto, $N \sim (-1,1)$); (c) tamaño del DIF indicado por la medida de área sin signo (Raju, 1998): DIF pequeño (USA = 0.30) y grande (USA = 0.60) para el caso de DIF unidireccional; pequeño, moderado, grande y muy grande (USA = 0.43, 0.64, 0.85 y 1.06, respectivamente) para el DIF no unidireccional.

Los resultados muestran que el método IRT LRT tuvo un adecuado control del error Tipo I en todas las situaciones. Las tasas de error Tipo I promedio fueron de 0.05 y 0.06 para tamaños de muestra iguales y diferentes (500-500 y 900-100), independientemente de la presencia de impacto. Por el contrario, el error Tipo I en RL

estaba inflado en las condiciones de impacto, especialmente en la condición de muestras iguales (500-500). Los autores argumentan que RL podría mejorar si se la utilizara como variable de igualación la puntuación corregida que se utiliza en CATSIB. En relación con CATSIB, Lei et al. (2006) concluyen que incluir la corrección por regresión atenúa el problema de la inflación en la tasa de error Tipo I, que se ve más afectada cuando el tamaño de las muestras es distinto, especialmente en presencia de impacto. Ese peor rendimiento parece asociarse a la presencia de celdas con pocos evaluados en las tablas de contingencia. Los autores encuentran que el problema se atenúa reduciendo el número de intervalos de habilidad.

Una vez eliminados del análisis de potencia en las condiciones en las que la tasa de error Tipo I fue inaceptable, el funcionamiento de los métodos parece estar relacionado con el impacto, las características del ítem (la potencia fue más alta cuando el ítem era más discriminativo), y la igualdad de los tamaños muestrales.

En relación al DIF unidireccional CATSIB, tiene mayor potencia que IRT LRT y RL para detectar DIF pequeño en las condiciones en las que no se sobreestima la tasa de error tipo I. En general, IRT LRT muestra mayor potencia que RL para detectar el DIF unidireccional.

En relación al DIF no unidireccional, IRT LRT y RL mostraron mayor potencia cuando los ítems eran más discriminativos tanto para el grupo focal ($a_F \geq 1.68$) como para el grupo de referencia ($a_R \geq 0.47$), CATSIB, mostró una potencia baja para identificar DIF, sobre todo para ítems de dificultad media (b=0), ya que en estos ítems se produce un efecto de cancelación del DIF a través del nivel de rasgo. IRT LRT mostró menor rendimiento en la detección del DIF no-unidireccional para ítems muy difíciles (b = 1.5) y poco discriminativos ($a_F \leq 1.03$ y $a_R \leq 0.50$) cuando el tamaño de las muestras era distinto, lo que los autores atribuyen a los altos errores típicos de estimación de los parámetros en esas condiciones. Si se comparan IRT LRT respecto a RL, que son métodos capaces de detectar DIF no unidireccional, el primero presenta mejores resultados cuando el tamaño de muestra fue 900–100 especialmente en la condición de impacto, mientras que RL tuvo mejores resultados que IRT LRT cuando el tamaño de las muestras fue 500–500 en la condición sin impacto.

Finalmente, si se observan ambas formas de DIF, CATSIB presentó mejores resultados en la condición de DIF unidireccional respecto del DIF no unidireccional (0.73 promedio vs. 0.42 respectivamente), mientras que IRT LRT y RL tuvieron

potencias similares en ambas condiciones (para IRT LRT fue 0.65 vs. 0.59; mientras que para RL fue de 0.62 vs. 0.59).

3.3. Conclusiones

Los resultados de los métodos adaptados para el análisis del DIF en TAIs permiten concluir que los resultados obtenidos por IRT LRT son prometedores, ya que es el único método que muestra tasas de error Tipo I aceptables en todas las condiciones. Esto es muy importante cuando se evalúan ítems operativos donde la eliminación de los ítems implica un mayor coste y por lo tanto se desea disminuir al máximo el número de falsos positivos.

Una ventaja adicional es que IRT LRT no requiere tomar decisiones que son importantes y que pueden afectar a la efectividad de la técnica. Por ejemplo, en relación a los métodos clásicos no se ha establecido claramente la forma de construir el número de intervalos que se deben emplear para agrupar a los examinados en la variable de igualación. En los métodos basados en tablas de contingencia (MH, STD) los evaluados con un rasgo estimado que se encuentren en un intervalo de ya sea una o dos unidades se consideran como parte del mismo nivel de rasgo. Esta forma de construir los intervalos parece funcionar adecuadamente para el análisis de ítems operativos, pero no así para el análisis de ítems pretest donde no existe restricción del rasgo y se produce una dispersión en las tablas construidas con estos criterios (Zwick et al., 1994b). En el análisis de ítems pretest con CATSIB, Lei et al. (2006) parten de 20 intervalos pero finalmente se ven obligados a reducir (en algunas condiciones) el número de intervalos a 10 para reducir la inflación en la tasa de error Tipo I que se produce en presencia de impacto. Los puntos de corte para los intervalos eran arbitrariamente elegidos considerando las frecuencias esperadas en el grupo total (intervalos más pequeños en torno a la media de la distribución e intervalos más grandes hacia los extremos), pero aún no se sabe si dichos criterios deben cambiarse cuando se analizen ítems aplicados adaptativamente.

Otra decisión importante tiene que ver con si se debe o no establecer la corrección por regresión según se incluya o no el ítem en el cálculo de la variable de igualación. Para evitar la distorsión que introduce la presencia de impacto en los estadísticos de DIF se han seguido dos estrategias alternativas:

1) Si el test es largo o los ítems siguen el modelo de Rasch, se incluye la puntuación del ítem en el cálculo de la variable de igualación. Por ejemplo, esta es la estrategia que se sigue en los métodos MH, STD y RL tradicionales. Zwick et al. (1993) proponen que cuando se analice el DIF de un ítem operativo se considere como variable de igualación W_0, que se basa en la estimación del rasgo incluyendo al ítem. Para el análisis de ítems pretest proponen varias alternativas que siguen la lógica de añadir el ítem (Zwick et al., 1993, pag. 52).
2) Otro procedimiento para evitar el sesgo es la corrección por regresión propuesta por Shealy y Stout (1993b). Esta corrección ha sido propuesta para el análisis del DIF en ítems pretest (Nandakumar y Roussos, 2004). En ese caso, $\hat{\theta}_{TAI}$ hace el papel de variable de igualación calculada sin incluir el ítem pretest.

Ambas estrategias se conciben como estrategias alternativas. Es decir, si se incluye el ítem no debería aplicarse la corrección por regresión. Así, diversos autores sugieren la aplicación de la corrección sin incluir el ítem a técnicas clásicas distintas de CATSIB como MH o RL (DeMars, 2009; Monahan y Ankenman, 2010). Walker, et al. (2001) sugieren que si se evalúa un ítem operativo que usualmente participa en la estimación de la habilidad, $\hat{\theta}_{TAI}$ no debería aplicarse la corrección por regresión. En el contexto de los tests fijos ellos encuentran un aumento de la tasa de falsos positivos. Sin embargo, su estudio no se desarrolla en el contexto de los TAIs y se debe ser cauto a la hora de generalizar los resultados obtenidos en un test fijo a los obtenidos en un TAI.

Finalmente, aunque el CATSIB ha mostrado resultados prometedores en relación al IRT LRT en la detección del DIF unidireccional, no se puede prever resultados similares a los del estudio de Lei et al. (2006) en todos los casos, ya que mientras que para CATSIB y RL utilizaban la $\hat{\theta}_{TAI}$ estimada en el TAI, para aplicar IRT LRT se utilizaba $\hat{\theta}_{TAI}$ para la imputación de las respuestas pero se trabajaba sólo con 84 ítems de anclaje (de los 360 posibles ítems del banco).

Por otro lado, la sugerencia de Nandakumar y Roussos (2004) de que el método CATSIB puede fácilmente generalizarse a la evaluación de ítems operativos, debe tomarse con reserva. Por ejemplo, los estudios de Zwick et al. (1993, 1994b) demuestran que $\hat{\Delta}_{MH}$ funciona mejor para ítems operativos que para ítems prestest, mientras que para STD ocurre lo contrario. Esto ocurre porque el error típico de STD se incrementa a medida que la proporción de ítems se aproxima a 0.5, algo similar puede

ocurrir con CATSIB que tiene una métrica similar a STD. Además, Walker et al. (2001) encuentran que las tasas de falsos positivos en CATSIB pueden verse seriamente afectadas en presencia de contaminación. Esta es una situación esperable en un contexto de revisión del banco por lo que debería evaluarse como de robusto es CATSIB en esas condiciones.

A pesar de todo lo anterior, la aplicación hoy por hoy de IRT LRT sigue presentando.dificultades. En primer lugar, se requiere resolver algunos de los problemas encontrados por Lei et al. (2006) en la aplicación del *software* IRTLRTDIF (Thissen, 2001). El número de ítems que se pueden analizar es limitado en la versión actual del programa. Esto resulta insuficiente considerando que un banco realista de ítems tendrá usualmente más de 200 ítems. Además, el *software* actual resulta ineficiente pues requiere estimar los parámetros de todo el banco de ítems cada vez que se analiza el DIF de un ítem. En segundo lugar, la propuesta de Lei et al. (2006) conlleva la imputación de respuestas, desconociéndose el nivel de distorsión que puede producirse cuando el nivel de rasgo no se estima con precisión. Finalmente, se desconoce el funcionamiento de IRT LRT en la detección del DIF en ítems operativos, en una situación en la que la restricción de rango en el nivel de rasgo de las muestras a las que se aplica el ítem estudiado es mayor.

En los dos siguientes capítulos se propone y contrasta una modificación del IRT LRT tradicional, basada en los procedimientos de calibración con parámetros fijos que se emplean en el análisis de ítems pretest (Kim, 2006). Primero se describen estos procedimientos, así como sus ventajas en este contexto. En definitiva, nuestra propuesta resuelve los problemas asociados a la estrategia de imputación. En el Estudio 1 se compara empíricamente nuestra propuesta (varías variantes) con el procedimiento de imputación. Este primer análisis se realiza en el contexto de detección del DIF en ítems pretest.

En el Estudio 2 se compara el nuevo procedimiento con la propuesta clásica que ha proporcionado mejores resultados empíricos en los trabajos previos: CATSIB. Ni CATSIB ni IRT LRT han sido analizados en el contexto de la detección del DIF en ítems operativos, por lo que nuestro trabajo permite contrastar la viabilidad de estas técnicas para el estudio del DIF.

Capítulo 4

Calibración online y análisis del DIF en ítems pretest

4.1. Introducción

Los procedimientos de calibración online se utilizan para renovar o ampliar el banco de ítems en test adaptativos informatizados (TAIs). La particularidad de un diseño de calibración online es que los evaluados reciben los ítems nuevos (pretest) durante una sesión adaptativa informatizada. El objetivo principal es estimar los parámetros de esos ítems con la ayuda de la información que proporcionan las respuestas al TAI. Un objetivo adicional es garantizar que los parámetros de los ítems pretest se sitúan en la misma escala métrica que la escala de los ítems operativos. De esta forma se asegura que los tests construidos a partir de los nuevos ítems dan lugar a estimaciones comparables del nivel de rasgo (Wainer y Mislevy, 1990; Stocking, 1988).

Una ventaja del diseño de calibración online es que los ítems pretest son aplicados a muestras representativas y motivadas en una situación de evaluación real, generalmente en el formato final en el que se aplicarán cuando se conviertan en operativos (Parshall, 1998). Sin embargo, la calibración online resulta compleja, puesto que: (a) se obtienen matrices de respuestas dispersas, en las que algunos de los ítems operativos se aplican muy pocas veces o no se aplican nunca en la muestra de calibración (Hsu, Thompson y Chen, 1998); y (b) los ítems operativos son respondidos por muestras de rango restringido en los niveles de rasgo. Todo ello puede dar lugar a estimaciones imprecisas de los parámetros y/o a problemas de convergencia en los métodos de estimación (Ban, Hanson, Wang, Yi y Harris, 2001).

Dos formas de calibración pueden aplicarse a este tipo de diseños: 1) la calibración independiente con un procedimiento posterior de anclaje; o 2) la calibración con parámetros fijos (CPF). En la calibración independiente se calibran tanto los ítems pretest como los ítems operativos sin considerar la escala existente. Por ello, se hace necesario situar los parámetros de los ítems pretest en la escala de los ítems operativos. La ecuación de transformación se establece mediante algún procedimiento de anclaje usando los ítems operativos como ítems de anclaje (p. ej., Stocking, 1988). Este método supone un tiempo de cálculo considerable, dado que para la calibración de cada ítem pretest se requiere la reestimación de los parámetros de todos los ítems operativos.

En el caso de la CPF, se fijan los parámetros de los ítems operativos a sus valores previamente estimados y se calibran únicamente los ítems nuevos. De este modo, la escala métrica ya está fijada. Los procedimientos de CPF han mostrado mayor eficacia en la estimación de los parámetros que los métodos de calibración independiente (Ban et al., 2001; Kim, 2006).

Además de ser calibrado, cada ítem pretest debe pasar por un proceso de revisión que incluye una serie de análisis estadísticos para determinar sus propiedades psicométricas. Sólo si el ítem muestra propiedades psicométricas favorables, éste pasa a formar parte del banco. Uno de los análisis de validez más importantes consiste en estudiar el funcionamiento diferencial del ítem o DIF. Como ya se ha descrito previamente, el análisis del DIF permite determinar si las funciones de respuesta del ítem difieren significativamente a través de los grupos.

Thissen, Steinberg y Wainer (1988) propusieron el uso de la prueba de razón de verosimilitudes de la TRI (IRT LRT) para el análisis del DIF en TAIs. Sin embargo, hasta el momento el método IRT LRT se ha implementado únicamente con procedimientos de calibración independiente. La extensión multigrupo de los procedimientos de CPF puede ofrecer un tratamiento unificado de los procesos de calibración y análisis del DIF en ítems pretest y resolver algunos de los problemas (pérdida de la métrica; tiempo de cálculo) que se han encontrado en la aplicación directa del test IRT LR.

Antes de poder introducir con detalle los métodos de CPF multigrupo, resulta necesario introducir otras técnicas de estimación de parámetros. Por ello, en los apartados siguientes se describe la aplicación del algoritmo EM general en la estimación por Máxima Verosimilitud Marginal, se describe la aplicación al caso multigrupo, el modelo de calibración independiente y los diversos modos de

implementación para la calibración con parámetros fijos. Finalmente, se describe la aplicación al análisis del DIF y cuáles son los resultados esperados en este contexto.

4.2. Aplicación del algoritmo EM en la estimación por Máxima Verosimilitud Marginal (MML)

El método de Máxima Verosimilitud Marginal (MML) es probablemente el método más ampliamente implementado para la estimación de los parámetros de los ítems ($\boldsymbol{\Delta}$). Esta técnica fue desarrollada por Bock y Aitkin (1981) para resolver el problema de la inconsistencia en la estimación de $\boldsymbol{\Delta}$, debido a que $f(\mathbf{U})$, la función de verosimilitud de los datos ($\mathbf{U}$), depende también de los parámetros de los sujetos ($\boldsymbol{\theta}$): $f(\mathbf{U} \mid \boldsymbol{\Delta}, \boldsymbol{\theta})$.

Bock y Aitkin (1981) proponen añadir el supuesto de que θ sigue en la población una distribución conocida o que puede ser estimada. Así, en el método MML se obtienen los parámetros $\boldsymbol{\Delta}$ y $\boldsymbol{\pi}$ que maximizan la función de verosimilitud marginal:

$$f(\mathbf{U} \mid \boldsymbol{\Delta}, \boldsymbol{\pi}) = \int f(\mathbf{U} \mid \boldsymbol{\Delta}, \theta) f(\theta \mid \boldsymbol{\pi}) d\theta, \tag{4.1}$$

donde $f(\theta \mid \boldsymbol{\pi})$ indica la distribución del rasgo en la población y $\boldsymbol{\pi}$ es el vector de parámetros que determinan la distribución de θ (p.ej., si θ sigue una distribución normal, $\boldsymbol{\pi}$ sería el vector de parámetros que contiene la media, μ, y la desviación típica de la distribución, σ).

De esta manera, al integrar sobre la distribución del rasgo, se elimina θ de la función de verosimilitud. Como contrapartida, se asume una distribución del rasgo en la población (Baker y Kim, 2004; Harwell y Baker, 1991).

El método MML incorpora el algoritmo EM (Esperanza-Maximización) para la estimación de los parámetros. En el algoritmo EM se utiliza la verosimilitud de los datos completos [$f(\mathbf{U}, \boldsymbol{\theta} \mid \boldsymbol{\Delta}, \boldsymbol{\pi})$] para encontrar los valores de los parámetros (de los ítems y de la distribución del rasgo) que maximizan la verosimilitud de los datos observados [$f(\mathbf{U} \mid \boldsymbol{\Delta}, \boldsymbol{\pi})$] (Woodruff y Hanson, 1996). En concreto, se sigue un procedimiento iterativo en el que, en cada ciclo, los estimadores de los parámetros maximizan el valor esperado de $f(\mathbf{U}, \theta \mid \boldsymbol{\Delta}, \boldsymbol{\pi})$ dadas las respuestas observadas y dados los parámetros provisionales obtenidos en el ciclo anterior ($\hat{\boldsymbol{\Delta}}^{(s-1)}, \hat{\boldsymbol{\pi}}^{(s-1)}$). En cada ciclo s

los parámetros se obtienen como:

$$\left(\hat{\boldsymbol{\Delta}}^{(s)},\hat{\boldsymbol{\pi}}^{(s)}\right)=\arg\max_{\boldsymbol{\Delta},\boldsymbol{\pi}} E_{\Theta}(\log f(\mathbf{U},\theta\,|\,\boldsymbol{\Delta},\boldsymbol{\pi})\,|\,\mathbf{U},\hat{\boldsymbol{\Delta}}^{(s-1)},\hat{\boldsymbol{\pi}}^{(s-1)}). \quad (4.2)$$

Las iteraciones finalizan cuando las diferencias entre los parámetros de ciclos sucesivos están por debajo de un valor arbitrariamente pequeño.

La aplicación del algoritmo EM para la estimación MML de los parámetros de los ítems en un test fijo aplicado a una única muestra se ha descrito en diversos trabajos (e.g., Bock y Aitkin, 1981; Mislevy y Bock, 1985; Woodruff y Hanson, 1996). En el siguiente apartado se introduce la notación para describir la aplicación del algoritmo EM al estudio del DIF con matrices de datos incompletas (p.ej., TAIs) y se describe dicho algoritmo para el caso más general, en el que se estiman los parámetros de todos los ítems.

4.3. La aplicación del algoritmo EM a un modelo multigrupo

Durante la sesión de evaluación, cada evaluado i_g responde al TAI y a un ítem pretest[2]. i_g se refiere a la persona i del grupo g ($i_g = 1, \ldots, N_g$, donde N_g es el número de evaluados del grupo g); g indica el grupo de pertenencia ($g = 1, 2$), donde 1 indica el grupo de referencia y 2 el grupo focal. Las respuestas de los evaluados pueden organizarse en una matriz de datos incompleta de dimensiones $N \times J$, donde N es el número total de evaluados ($N = N_1 + N_2$) y J es el número de ítems total, incluyendo los ítems operativos del banco y el ítem pretest ($J = J_{\text{ope}} + 1$).

Adaptando la notación de Mislevy y Wu (1996), los datos para dicho evaluado se describen por:

- $\mathbf{u}_{i_g} = (u_{i_g 1},\ldots,u_{i_g J})$, es la matriz que se genera de una aplicación adaptativa y contiene las respuestas a todos los ítems, tanto los administrados como los no administrados, puede escribirse como $\mathbf{u}_{i_g} = (\mathbf{u}_{i_g(obs)}, \mathbf{u}_{i_g(miss)})$. para distinguir entre los elementos observados y no observados de $\mathbf{u}_{i_g}$. Donde los $\mathbf{u}_{i_g(obs)}$ pueden ser 1

[2] Para simplificar la exposición se asume que los evaluados responden a un único ítem pretest.

o 0 para indicar si la respuesta es correcta o incorrecta y $\mathbf{u}_{i_g(miss)}$) puede tomar cualquier valor (i.e. 9). También puede escribirse $\mathbf{u}_{i_g} = (\mathbf{u}_{i_g(ope)}, \mathbf{u}_{i_g(pre)})$, para distinguir entre las respuestas a los ítems operativos y las respuestas a los ítems pretest.

- $\mathbf{m}_{i_g} = (m_{i_g 1}, ..., m_{i_g J})$; cada elemento $m_{i_g j}$ de este vector indica si el ítem *j* ha sido aplicado o no ($m_{i_g j}$ = 0, 1). $m_{i_g j} = 0$ indica que el ítem no se ha aplicado ($u_{i_g j}$ no se ha observado); $m_{i_g j} = 1$ indica que el ítem j se ha aplicado ($u_{i_g j}$ se ha observado).
- θ_{i_g} indica el nivel de rasgo del evaluado i_g, que es desconocido.
- $\boldsymbol{\Delta}$ es una matriz que contiene los parámetros de los ítems y $\boldsymbol{\pi}$ es un vector que contiene los parámetros que definen la distribución de θ. Los parámetros se pueden diferenciar para cada grupo:

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2) \tag{4.3}$$

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) \tag{4.4}$$

Además, los parámetros de los ítems se pueden diferenciar según el subconjunto de ítems que se trate:

$$\boldsymbol{\Delta}_g = (\boldsymbol{\Delta}_{g(ope)}, \boldsymbol{\Delta}_{g(pre)}) \tag{4.5}$$

Cuando se aplica un TAI la probabilidad del patrón de respuestas observadas para un evaluado, asumiendo que las respuestas a los ítems son independientes, es[3]:

$$f(\mathbf{u}_{i_g(obs)} \left| \theta_{i_g}, \boldsymbol{\Delta}_g \right.) = \prod_{j=1}^{J} P_j(u_{i_g j} \mid \theta_{i_g})^{m_{i_g j}} \tag{4.6}$$

donde $P_j(u_{i_g j} \mid \theta_{i_g})$ indica la probabilidad de la respuesta al ítem *j*, que depende del modelo de TRI con el que se trabaje. La función de log-verosimilitud de $\mathbf{U}_{(obs)}$ asumiendo una distribución de θ (discretizada) para cada grupo g:

[3] Nótese que se presenta $f(\mathbf{u}_{i_g(obs)} \mid \theta_{i_g}, \boldsymbol{\Delta}_g)$ cuando lo correcto sería $f(\mathbf{u}_{i_g(obs)}, \mathbf{m}_{i_g} \mid \theta_{i_g}, \boldsymbol{\Delta}_g)$. En el Anexo se detallan las razones.

$$\log f(\mathbf{U}_{(obs)}|\boldsymbol{\Delta},\boldsymbol{\pi}) = \log\left[\prod_{g=1}^{2}\prod_{i_g=1}^{N_g}\sum_{k_g}^{K_g}\left\{f(\mathbf{u}_{i_g}|\boldsymbol{\Delta}_g,\theta_{k_g})f(\theta_{k_g}|\boldsymbol{\pi}_g)\right\}\right] \tag{4.7}$$

donde k_g se refiere al punto de cuadratura k dentro del grupo g con nivel de rasgo θ_{k_g} (k_g = 1,…,K_g, siendo K_g el número de puntos de cuadratura). $f(\theta_{k_g}|\boldsymbol{\pi}_g)$ se denomina distribución previa.

Como ya se ha mencionado, el algoritmo EM se basa en el hecho de que maximizar la verosimilitud de $\mathbf{U}_{(obs)}$ es equivalente a maximizar el valor esperado condicional de la verosimilitud conjunta de $\mathbf{U}_{(obs)}$ y $\boldsymbol{\theta}$ (iterativamente):

$$\left(\hat{\boldsymbol{\Delta}}^{(s)},\hat{\boldsymbol{\pi}}^{(s)}\right) = \arg\max_{\boldsymbol{\Delta},\boldsymbol{\pi}} \mathrm{E}_{\Theta}\left\{\log\left[f(\mathbf{U}_{(obs)},\theta|\boldsymbol{\Delta},\boldsymbol{\pi})\right]\middle|\mathbf{U}_{(obs)},\hat{\boldsymbol{\Delta}}^{(s-1)},\hat{\boldsymbol{\pi}}^{(s-1)}\right\} \tag{4.8}$$

donde el valor esperado es condicional a los datos observados y a los valores "provisionales" (i. e., obtenidos en el paso s-1) de los parámetros de los ítems (Kim, 2006; McLachlan y Krishnan, 1997). Desarrollando la ecuación anterior:

$$\left(\hat{\boldsymbol{\Delta}}^{(s)},\hat{\boldsymbol{\pi}}^{(s)}\right) = \arg\max_{\boldsymbol{\Delta},\boldsymbol{\pi}} \mathrm{E}_{\Theta}\left\{\log\left[\prod_{g=1}^{2}\prod_{i_g=1}^{N_g}\left\{f(\mathbf{u}_{i_g(obs)}|\boldsymbol{\Delta}_g,\theta)f(\theta|\boldsymbol{\pi}_g)\right\}\right]\middle|\mathbf{U}_{(obs)},\hat{\boldsymbol{\Delta}}^{(s-1)},\hat{\boldsymbol{\pi}}^{(s-1)}\right\} =$$

$$= \arg\max_{\boldsymbol{\Delta},\boldsymbol{\pi}} \mathrm{E}_{\Theta}\left\{\left[\sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\log\left[f(\mathbf{u}_{i_g(obs)}|\boldsymbol{\Delta}_g,\theta)f(\theta|\boldsymbol{\pi}_g)\right]\right]\middle|\mathbf{U}_{(obs)},\hat{\boldsymbol{\Delta}}^{(s-1)},\hat{\boldsymbol{\pi}}^{(s-1)}\right\} =$$

$$= \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\mathrm{E}_{\Theta}\left\{\log\left[f(\mathbf{u}_{i_g(obs)}|\boldsymbol{\Delta}_g,\theta)f(\theta|\boldsymbol{\pi}_g)\right]\middle|\mathbf{u}_{i_g(obs)},\hat{\boldsymbol{\Delta}}_g^{(s-1)},\hat{\boldsymbol{\pi}}_g^{(s-1)}\right\} =$$

$$= \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\mathbf{u}_{i_g(obs)}|\boldsymbol{\Delta}_g,\theta_{k_g})f(\theta_{k_g}|\boldsymbol{\pi}_g)\right]f(\theta_{k_g}\,|\,\mathbf{u}_{i_g(obs)},\hat{\boldsymbol{\Delta}}_g^{(s-1)},\hat{\boldsymbol{\pi}}_g^{(s-1)})\right\} \tag{4.9}$$

Por otro lado, la ecuación a maximizar se puede separar en dos términos:

$$\left(\hat{\boldsymbol{\Delta}}^{(s)},\hat{\boldsymbol{\pi}}^{(s)}\right) = \arg\max_{\boldsymbol{\Delta},\boldsymbol{\pi}}\left(\phi(\boldsymbol{\Delta})+\psi(\boldsymbol{\pi})\right) \tag{4.10}$$

donde

$$\phi(\boldsymbol{\Delta}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\mathbf{u}_{i_g(obs)}|\boldsymbol{\Delta}_g,\theta_{k_g})\right]f(\theta_{k_g}\,|\,\mathbf{u}_{i_g(obs)},\hat{\boldsymbol{\Delta}}_g^{(s-1)},\hat{\boldsymbol{\pi}}_g^{(s-1)})\right\} \tag{4.11}$$

y

$$\psi(\boldsymbol{\pi}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\theta_{k_g}|\boldsymbol{\pi}_g)\right]f(\theta_{k_g}\,|\,\mathbf{u}_{i_g(obs)},\hat{\boldsymbol{\Delta}}_g^{(s-1)},\hat{\boldsymbol{\pi}}_g^{(s-1)})\right\} \tag{4.12}$$

Con esta separación, el primer término $\phi(\boldsymbol{\Delta})$ depende únicamente de $\boldsymbol{\Delta}$, mientras que $\psi(\boldsymbol{\pi})$ depende de $\boldsymbol{\pi}$. El término $f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)}, \hat{\boldsymbol{\Delta}}_g^{(s-1)}, \hat{\boldsymbol{\pi}}_g^{(s-1)})$ se denomina verosimilitud posterior y se calcula a partir de las respuestas observadas a los ítems $\mathbf{u}_{i_g(obs)}$ y los valores provisionales de los parámetros obtenidos en el ciclo *s* - 1:

$$f(\theta_{k_g} \mid \mathbf{u}_{i_g}, \hat{\boldsymbol{\Delta}}_g^{(s-1)}, \hat{\boldsymbol{\pi}}_g^{(s-1)}) = \frac{f(\mathbf{u}_{i_g(obs)} \mid \theta_{k_g}, \hat{\boldsymbol{\Delta}}_g^{(s-1)}) f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}_{k_g}^{(s-1)})}{\sum_{k'_g=1}^{K_g} f(\mathbf{u}_{i_g(obs)} \mid \theta_{k'_g}, \hat{\boldsymbol{\Delta}}_g^{(s-1)}) f(\theta_{k'_g} \mid \hat{\boldsymbol{\pi}}_{k_g}^{(s-1)})} \tag{4.13}$$

donde $f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}_{k_g}^{(s-1)})$ es la distribución previa de θ_{k_g}.

El algoritmo EM se ejecuta en varios ciclos. En cada ciclo s se ejecutan dos pasos (E y M):

- *Paso E (Esperanza):* Se calculan las verosimilitudes posteriores $f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)}, \hat{\boldsymbol{\Delta}}_g^{(s-1)}, \hat{\boldsymbol{\pi}}_g^{(s-1)})$. En el primer ciclo (*s* = 1) se parte de unos valores iniciales para los parámetros $(\boldsymbol{\Delta}_g^{(0)}, \boldsymbol{\pi}_g^{(0)})$. $f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}_{k_g}^{(0)})$ se denomina distribución inicial.
- *Paso M (Maximización):* Se calculan los parámetros estimados $\hat{\boldsymbol{\Delta}}_g^{(s)}$ que maximizan la ecuación $\phi(\boldsymbol{\Delta})$ y los parámetros estimados $\hat{\boldsymbol{\pi}}_g^{(s)}$ que maximizan la ecuación $\psi(\boldsymbol{\pi})$:

$$\hat{\boldsymbol{\Delta}}^{(s)} = \arg\max_{\boldsymbol{\Delta}} \phi(\boldsymbol{\Delta}) \tag{4.14}$$

$$\hat{\boldsymbol{\pi}}^{(s)} = \arg\max_{\boldsymbol{\pi}} \left(\psi(\boldsymbol{\pi})\right) \tag{4.15}$$

Las estimaciones de ambos conjuntos de parámetros se realizan por técnicas iterativas (p.ej., Newton-Raphson). La estimación modal bayesiana se puede obtener si se añade el logaritmo de la distribución previa de $\boldsymbol{\Delta}_g$ a $\phi(\boldsymbol{\Delta}_g)$ y/o el logaritmo de la distribución previa de $\boldsymbol{\pi}_g$ a $\psi(\boldsymbol{\pi}_g)$.

4.4. Restricciones para la identificación en un modelo multigrupo

Es necesario imponer algunas restricciones para que el modelo esté identificado.

Como se ha mencionado, en los modelos de TRI la localización y la escala de la variable θ no es única y como consecuencia se pueden plantear diversas transformaciones lineales de la misma sin que varíen las características del modelo. Para la estimación es necesario resolver la indeterminación de la escala. Para ello, usualmente:

- Se especifica que θ tiene media 0 y desviación típica 1 en el grupo de referencia, $\boldsymbol{\pi}_1 = (0,1)$.
- Se especifican los ítems de anclaje. En nuestro caso se especifica que los parámetros de los ítems operativos son iguales en ambos grupos ($\boldsymbol{\Delta}_{1(ope)} = \boldsymbol{\Delta}_{2(ope)}$). De esta manera, los parámetros de distribución en el grupo focal ($\boldsymbol{\pi}_2$) pueden estimarse libremente.

4.5. Calibración con Parámetros Fijos (CPF)

En la primera parte de la exposición se ha presentado el caso multigrupo en el que se estiman los parámetros tanto de los ítems operativos como del ítem pretest. El procedimiento convencional tiene dos limitaciones importantes:

- La estimación de parámetros en ítems operativos que han sido escasamente aplicados puede resultar problemática.
- Si se especifica $\boldsymbol{\pi}_1 = (0,1)$, los parámetros de los ítems (operativos y pretest) y de distribución de los grupos (referencia y focal) se obtendrán en una escala métrica distinta a la métrica original.

Para resolver estos inconvenientes, Kim (2006) propone el uso de métodos de calibración con parámetros fijos, en los que se fijan los parámetros de los ítems operativos a sus estimaciones conocidas. De esta forma sólo se requiere estimar los parámetros de los ítems pretest ($\boldsymbol{\Delta}_{(pre)}$) y se pueden estimar los parámetros de distribución de ambos grupos, con la ventaja de que todos los parámetros se estiman automáticamente en la métrica original de los ítems operativos.

Existen distintas alternativas para aplicar un método CPF en función de cómo se

concrete el algoritmo EM. Esto es, depende de si la información de los nuevos ítems pretest se utiliza en el algoritmo para actualizar la distribución previa y/o para calcular la verosimilitud posterior. En función de esto, dos variables permiten clasificar los distintos procedimientos de CPF: 1) el número de ciclos EM utilizados para la estimación; y 2) el número de veces que se actualiza la distribución previa. Las combinaciones posibles de ambas variables se representan en la Tabla 4.1.

Tabla 4.1. Marco para la clasificación de los métodos de CPF (Kim, 2006).

		Número de ciclos EM	
		Un ciclo (OEM)	Múltiples ciclos (MEM)
Número de actualizaciones de la distribución previa	Ninguna (NWU)	NWU-OEM	NWU-MEM
	Una (OWU)	OWU-OEM	OWU-MEM
	Múltiples (MWU)	- - -	MWU-MEM

Cuando nos referimos al número de ciclos EM, podemos encontrar:

- *Un ciclo EM (OEM).* En éste procedimiento en el único paso E se calcula la verosimilitud posterior empleando las respuestas a los ítems operativos $\mathbf{u}_{i(obs)(ope)_g}$ y los parámetros de los mismos $\boldsymbol{\Delta}_{g(ope)}$. Sea $f(\theta_{k_g} \mid \boldsymbol{\pi}^*_g)$ la distribución previa. La verosimilitud posterior se calcula como:

$$f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)(ope)}, \boldsymbol{\Delta}_{g(ope)}, \hat{\boldsymbol{\pi}}^*_g, \mathbf{m}_{i_g}) = \frac{f(\mathbf{u}_{i_g(obs)(ope)} \mid \theta_{k_g}, \boldsymbol{\Delta}_{g(ope)}) f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}^*_{k_g})}{\sum_{k'_g=1}^{K_g} f(\mathbf{u}_{i_g(obs)(ope)} \mid \theta_{k'_g}, \boldsymbol{\Delta}_{g(ope)}) f(\theta_{k'_g} \mid \hat{\boldsymbol{\pi}}^*_{k_g})} \quad (4.16)$$

  Puesto que los parámetros de los ítems operativos son fijos sólo es necesario un único ciclo.

- *Múltiples ciclos EM (MEM).* En éste procedimiento a partir del segundo paso E se calcula la verosimilitud posterior utilizando los datos y parámetros de todos los ítems, incluido el ítem pretest.

$$f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)}, \boldsymbol{\Delta}_{g(ope)}, \hat{\boldsymbol{\Delta}}^{(s-1)}_{g(pre)}, \hat{\boldsymbol{\pi}}^{*}_{g}) = \frac{f(\mathbf{u}_{i_g(obs)} \mid \theta_{k_g}, \boldsymbol{\Delta}_{g(ope)}, \hat{\boldsymbol{\Delta}}^{(s-1)}_{g(pre)}) f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}^{*}_{k_g})}{\sum_{k'_g=1}^{K_g} f(\mathbf{u}_{i_g(obs)} \mid \theta_{k'_g}, \boldsymbol{\Delta}_{g(ope)}, \hat{\boldsymbol{\Delta}}^{(s-1)}_{g(pre)}) f(\theta_{k'_g} \mid \hat{\boldsymbol{\pi}}^{*}_{k_g})} \quad (4.17)$$

Puesto que los parámetros del ítems pretest se actualizan iterativamente es necesaria una actualización en cada ciclo *s* (en cada ciclo *s*, se utilizan los parámetros de los ítems pretest estimados en el ciclo anterior).

Respecto al número de veces que se actualiza la distribución previa los métodos se pueden caracterizar por:

- *No actualizar la distribución previa* (No Prior Weights Updating, NWU). En este caso se asumen, a través de todo el proceso de estimación, distribuciones previas arbitrarias:

$$f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}^{*}_{k_g}) = f(\theta_{k_g} \mid \boldsymbol{\pi}^{(0)}_{g}) \quad (4.18)$$

- *Una sola actualización de la distribución previa* (One Prior Weights Updating, OWU). En este caso la distribución previa del rasgo se actualiza una única vez considerando los parámetros y las respuestas a los ítems operativos del TAI. Para ello se obtienen los parámetros $\hat{\boldsymbol{\pi}}^{(1)}_{g}$ que maximizan:

$$\boldsymbol{\pi}^{(1)}_{g} = \arg\max_{\boldsymbol{\pi}} \psi(\boldsymbol{\pi}) = \sum_{g=1}^{2} \sum_{i_g=1}^{N_g} \sum_{k_g=1}^{K_g} \left\{ \log\left[f(\theta_{k_g} \mid \boldsymbol{\pi}_{g}) \right] f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)(ope)}, \boldsymbol{\Delta}_{g(ope)}, \boldsymbol{\pi}^{(0)}_{g}) \right\} \quad (4.19)$$

Como distribución previa se utiliza:

$$f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}^{*}_{k_g}) = f(\theta_{k_g} \mid \boldsymbol{\pi}^{(1)}_{g}) \quad (4.20)$$

Puesto que los parámetros de los ítems operativos son fijos, sólo es necesaria una única actualización.

- *Múltiples actualizaciones de la distribución previa* (Multiple Prior Weights Updating, OWU). En este caso, la distribución previa para el ciclo *s* (*s* > 1) se calcula a partir de las respuestas de todos los ítems (operativos y pretest), $f(\theta_{k_g} \mid \boldsymbol{\pi}^{(s-1)}_{g})$. Para ello se obtienen los parámetros $\hat{\boldsymbol{\pi}}^{(s-1)}_{g}$ que maximizan:

$$\boldsymbol{\pi}_g^{(s-1)} = \arg\max_{\boldsymbol{\pi}} \psi(\boldsymbol{\pi}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\theta_{k_g}|\boldsymbol{\pi}_g)\right] f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)(ope)}, \boldsymbol{\Delta}_{g(ope)}, \hat{\boldsymbol{\Delta}}_{g(pre)}^{(s-2)}, \hat{\boldsymbol{\pi}}_g^{(s-2)})\right\} \tag{4.21}$$

Puesto que los parámetros del ítems pretest se actualizan iterativamente es necesaria una actualización en cada ciclo *s*. En cada ciclo *s*, se utiliza como distribución previa la estimada en el ciclo anterior:

$$f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}_{k_g}^{*}) = f(\theta_{k_g} \mid \boldsymbol{\pi}_g^{(s-1)}) \tag{4.22}$$

A continuación se describe cada uno de los métodos en mayor detalle.

4.5.1. No actualización de la distribución previa y un ciclo EM (NWU-OEM)

Este método sugerido por Wainer y Mislevy (1990) se realiza en un único ciclo. En el paso E se calcula la verosimilitud posterior de cada valor θ_{k_g} para cada evaluado i_g considerando únicamente las respuestas y los parámetros de los ítems operativos ($\mathbf{u}_{i_g(obs)(ope)}$ y $\boldsymbol{\Delta}_{g(ope)}$) y la distribución del rasgo inicial $\boldsymbol{\pi}_g^{(0)}$:

$$f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)(ope)}, \boldsymbol{\Delta}_{g(ope)}, \boldsymbol{\pi}_g^{(0)}) \tag{4.23}$$

y en el paso M se estiman los parámetros $\hat{\boldsymbol{\Delta}}_{(pre)}^{(1)}$ del ítem pretest que maximizan:

$$\phi(\boldsymbol{\Delta}_{(pre)}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\mathbf{u}_{i_g(obs)(pre)}|\boldsymbol{\Delta}_{g(pre)}, \theta_{k_g})\right] f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)(ope)}, \boldsymbol{\Delta}_{g(ope)}, \boldsymbol{\pi}_g^{(0)})\right\} \tag{4.24}$$

En este método la distribución previa no se actualiza y la verosimilitud posterior está afectada únicamente por los ítems operativos, que no se estiman, por lo que sólo es necesario un único ciclo. El método NWU-OEM no actualiza la distribución del rasgo, la cual se define, en la mayoría de los casos, de manera arbitraria. Por lo tanto, la falta de especificación de la distribución del rasgo previa podría afectar sistemáticamente la ejecución del método.

4.5.2 No actualización de la distribución previa y múltiples ciclos EM (NWU-MEM)

En este método, sugerido por Ban et al. (2001), el primer ciclo EM se realiza de la misma forma que el NWU-OEM. Al iniciar el segundo ciclo en el paso E, sin embargo, el NWU-MEM usa las respuestas tanto de los ítems operativos como de los ítems pretest para obtener los valores de la verosimilitud posterior de cada valor θ_{k_g}:

$$f(\theta_{k_g} \mid \mathbf{u}_{i(obs)_g}, \boldsymbol{\Delta}_{g(ope)}, \hat{\boldsymbol{\Delta}}_{g(pre)}^{(s-1)}, \boldsymbol{\pi}_g^{(0)}) \tag{4.25}$$

A partir del segundo paso M en adelante, los parámetros estimados de $\hat{\boldsymbol{\Delta}}_{(pre)}^{(s)}$ son los que maximizan:

$$\phi(\boldsymbol{\Delta}_{(pre)}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\mathbf{u}_{i_g(obs)(pre)}\middle|\boldsymbol{\Delta}_{g(pre)}, \theta_{k_g})\right] f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)}, \boldsymbol{\Delta}_{g(ope)}, \hat{\boldsymbol{\delta}}_{g(pre)}^{(s-1)}, \boldsymbol{\pi}_g^{(0)})\right\} \tag{4.26}$$

En este método la distribución previa no se actualiza pero la verosimilitud posterior está afectada por el ítem pretest, por lo que son necesarios varios ciclos EM. Este procedimiento presenta dos problemas: 1) al igual que en el anterior método la distribución previa de los grupos se define de manera arbitraria; 2) el incluir el ítem pretest en el cálculo de la verosimilitud posterior puede resultar negativo si el ítem pretest resulta mostrar un mal ajuste al modelo (Ban, et al., 2001).

4.5.3. Una sola actualización de la distribución previa y un ciclo EM (OWU-OEM)

Este método difiere del NWU-OEM en que antes de aplicar el ciclo EM se actualiza la distribución inicial. Para ello se obtienen los parámetros $\hat{\boldsymbol{\pi}}_g^{(1)}$ que maximizan:

$$\psi(\boldsymbol{\pi}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\theta_{k_g}\middle|\boldsymbol{\pi}_g)\right] f(\theta_{k_g} \mid \mathbf{u}_{i_g(obs)(ope)}, \boldsymbol{\Delta}_{g(ope)}, \boldsymbol{\pi}_g^{(0)})\right\} \tag{4.27}$$

La actualización de la distribución inicial es conveniente cuando no existe información específica sobre la distribución del rasgo para los evaluados de los grupos

en los que se está evaluando el DIF. En el paso M se estiman los parámetros $\hat{\boldsymbol{\Delta}}_{g(pre)}^{(1)}$ del ítem pretest que maximizan:

$$\phi(\boldsymbol{\Delta}_{(pre)}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\mathbf{u}_{i_g(obs)(pre)}\middle|\boldsymbol{\Delta}_{g(pre)},\theta_{k_g})\right]f(\theta_{k_g}\mid\mathbf{u}_{i_g(obs)(ope)},\boldsymbol{\Delta}_{g(ope)},\boldsymbol{\pi}_g^{(1)})\right\} \quad (4.28)$$

En este método la distribución previa que se actualiza y la verosimilitud posterior están afectadas únicamente por los ítems operativos, que no se estiman, por lo que sólo es necesario un único ciclo. Este procedimiento no es completamente eficiente porque no emplea la información de los nuevos ítems para la recuperación de la distribución y la estimación de los parámetros, por lo que puede encontrarse algún sesgo en la estimación.

4.5.4. Una sola actualización de la distribución previa y Múltiples ciclos EM (OWU-MEM)

En este método el primer ciclo EM se realiza de la misma forma que en el método NWU-OEM. De esta forma, se obtienen parámetros provisionales para los ítems pretest $\hat{\boldsymbol{\Delta}}_{g(pre)}^{(1)}$. Además, se obtienen los parámetros $\hat{\boldsymbol{\pi}}_g^{(1)}$ que maximizan (igual que en OWU-OEM):

$$\psi(\boldsymbol{\pi}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\theta_{k_g}\middle|\boldsymbol{\pi}_g)\right]f(\theta_{k_g}\mid\mathbf{u}_{i_g(obs)(ope)},\boldsymbol{\Delta}_{g(ope)},\boldsymbol{\pi}_g^{(0)})\right\} \quad (4.29)$$

A partir del segundo paso M en adelante, los parámetros estimados de $\hat{\boldsymbol{\Delta}}_{g(pre)}^{(s)}$ son los que maximizan:

$$\phi(\boldsymbol{\Delta}_{(pre)}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\mathbf{u}_{i_g(obs)(pre)}\middle|\boldsymbol{\Delta}_{g(pre)},\theta_{k_g})\right]f(\theta_{k_g}\mid\mathbf{u}_{i_g(obs)},\boldsymbol{\Delta}_{g(ope)},\hat{\boldsymbol{\Delta}}_{g(pre)}^{(s-1)},\boldsymbol{\pi}_g^{(1)})\right\} \quad (4.30)$$

En este método la distribución previa se actualiza una vez a partir de la información de los ítems operativos. Sin embargo, la verosimilitud posterior está afectada por el ítem pretest, por lo que son necesarios varios ciclos EM. De nuevo, el incluir el ítem pretest en el cálculo de la verosimilitud posterior puede resultar negativo

si el ítem pretest resulta mostrar un mal ajuste al modelo.

4.5.5. Múltiples actualizaciones de la distribución y múltiples ciclos EM (MWU-MEM)

En este método el primer ciclo EM se realiza de la misma forma que en el método NWU-OEM. De esta forma, se obtienen parámetros provisionales para los ítems pretest $\hat{\boldsymbol{\Delta}}_{g(pre)}^{(1)}$. Además, se obtienen los parámetros $\hat{\boldsymbol{\pi}}_g^{(1)}$ que maximizan (igual que en OWU-OEM):

$$\psi(\boldsymbol{\pi}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\theta_{k_g}|\boldsymbol{\pi}_g)\right]f(\theta_{k_g}\,|\,\mathbf{u}_{i_g(obs)(ope)},\boldsymbol{\Delta}_{g(ope)},\boldsymbol{\pi}_g^{(0)})\right\} \tag{4.31}$$

A partir del segundo paso M en adelante, se actualizan tanto los parámetros del ítem pretest ($\hat{\boldsymbol{\Delta}}_{g(pre)}^{(s)}$) como los parámetros de distribución ($\hat{\boldsymbol{\pi}}_g^{(s)}$). Para ello, en cada ciclo *s* se maximiza:

$$\phi(\boldsymbol{\Delta}_{(pre)}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\mathbf{u}_{i_g(obs)(pre)}\left|\boldsymbol{\Delta}_{g(pre)},\theta_{k_g}\right.)\right]f(\theta_{k_g}\,|\,\mathbf{u}_{i_g(obs)},\boldsymbol{\Delta}_{g(ope)},\hat{\boldsymbol{\Delta}}_{g(pre)}^{(s-1)},\boldsymbol{\pi}_g^{(s-1)}\right. \tag{4.32}$$

$$\psi(\boldsymbol{\pi}) = \sum_{g=1}^{2}\sum_{i_g=1}^{N_g}\sum_{k_g=1}^{K_g}\left\{\log\left[f(\theta_{k_g}|\boldsymbol{\pi}_g)\right]f(\theta_{k_g}\,|\,\mathbf{u}_{i_g(obs)},\boldsymbol{\Delta}_{g(ope)},\hat{\boldsymbol{\Delta}}_{g(pre)}^{(s-1)},\boldsymbol{\pi}_g^{(s-1)})\right\} \tag{4.33}$$

En este método el ítem pretest se utiliza doblemente, tanto para actualizar la distribución previa como para calcular la verosimilitud posterior, por lo que son necesarios varios ciclos EM. Una ventaja importante del método MWU-MEM es que utiliza la información adicional contenida en las respuestas al ítem pretest. Sin embargo, si el ítem pretest está desajustado esto puede afectar negativamente a los resultados de la calibración.

4.6. Estudios previos sobre la calibración de ítems con parámetros fijos

A pesar de las aparentes ventajas de los procedimientos de CPF pocos estudios los han evaluado y comparado con los procedimientos más tradicionales de calibración independiente (Ban, et al., 2001; Ban, Hanson, Yi, y Harris, 2002; Kim, 2006). Todos estos estudios se centran en el caso de un único grupo.

En Ban et al., (2001) el método NWU-OEM y su generalización a múltiples ciclos EM (NWU-MEM), se comparan con dos métodos de anclaje: el método A y el B de Stocking y con un procedimiento denominado "strong priors". En éste último método los valores de los parámetros de los ítems operativos se definen con distribuciones previas restrictivas: para la distribución previa de cada parámetro se fijaba una desviación típica muy pequeña y una media igual al valor estimado en la muestra de calibración inicial. En el diseño de calibración online se aplicaban de manera adaptativa, por máxima información, 30 ítems de un banco de 540 ítems con un modelo de 3PL. Al terminar el TAI se aplicaban 10 ítems pretest a todos los evaluados en nuevas muestras de calibración que podían ser de 300, 1000 y 3000 sujetos. En los resultados se observa que con el método NWU-MEM se mantenía la escala de los ítems operativos y se recuperaban los parámetros de los ítems pretest de manera adecuada (los valores de sesgo y RMSE fueron menores en todos los parámetros). El método B de Stocking obtuvo resultados similares a NWU-MEM. Sin embargo, para aplicar este procedimiento era necesario añadir un test de anclaje, lo que aumentaba considerablemente el tamaño del test. Con el resto de los métodos contrastados se producían sesgos en la recuperación de la escala métrica, por lo que los autores desaconsejaban su uso.

Ban, et al. (2002) extienden el estudio anterior comparando un mayor número de métodos y analizando el funcionamiento de éstos en un contexto en el que cada evaluado puede tomar únicamente un subconjunto de los ítems pretest. En este caso, existe una mayor dispersión en la matriz de respuestas a los ítems pretest. En el procedimiento se utilizaron las mismas condiciones de simulación que en el estudio previo con la diferencia de que se aplicaban distintos subconjuntos de 10 ítems (de un banco de 240 ítems pretest) en muestras de 250 evaluados. En el diseño los primeros 250 evaluados respondían a los primeros 10 ítems, los siguientes 250 respondían a los

ítems del 6 al 15 y los siguientes 250 respondían del 11 al 20; esta lógica continuaba hasta agotar la presentación del banco de ítems pretest, de forma que cada muestra compartía 5 ítems con la muestra previa y 5 con la muestra siguiente, la muestra final fue de 12000 evaluados. Los resultados de este estudio fueron consistentes con los encontrados en el estudio previo. El método NWU-MEM recuperaba mejor que el método B de Stocking la distribución del rasgo y los parámetros de los ítems. Asimismo, el rendimiento del método NWU-OEM fue peor que el encontrado en el estudio previo.

Finalmente, Kim (2006) comparó los cinco métodos de CPF en términos de la recuperación de la distribución del rasgo y de los parámetros de los ítems pretest.. En su estudio se simularon, con un modelo de 3PL, las respuestas a un test compuesto por 50 ítems operativos y pretest. Cuatro factores fueron considerados en la CPF: 1) Tamaño en la muestra previa de calibración (300 y 3000); 2) distribución del rasgo en la muestra de calibración ($N(0,1)$, $N(0.5,1.2)$, $N(0.5,1.4)$); 3) tamaño en la nueva muestra de calibración (300, 1000 y 3000); y 4) número de ítems operativos (10, 20, 30 ó 40, de los 50). El análisis de los resultados mostró que, independientemente del método de calibración, la precisión de los métodos era mayor cuanto mayores fueran los tamaños muestrales y/o el número de ítems operativos. Sin embargo, había importantes diferencias entre los métodos. En general, aquellos métodos que hacen uso de la información de los ítems pretest en el cálculo de la verosimilitud posterior (i.e., varios ciclos EM) dan lugar a una mejor recuperación de los parámetros de los ítems. Cuando la distribución de la muestra de calibración difería de la distribución previa inicial, la actualización de la distribución también era fundamental. En ese caso, sólo el MWU-MEM producía resultados ajustados; en el resto de los métodos se producía una contracción considerable de la distribución del rasgo estimada (subestimación de la media y de la desviación típica) y, por consiguiente, una peor recuperación de los parámetros de los ítems. Este efecto era mayor cuanto menor fuera el número de ítems operativos y/o si no se actualizaba la distribución del rasgo (la mayor precisión se encontraba para el método MWU y la peor para los métodos NWU).

En este estudio los métodos fueron evaluados en una situación en la que todos los ítems se ajustaban a un modelo de TRI, y todos los evaluados respondieron a todos los ítems de forma que los métodos se aplicaron a matrices de datos completas. Kim reconoce que estas dos características no permiten la generalización de los resultados a situaciones de calibración online.

4.7. Aplicación de los métodos de calibración con parámetros fijos al análisis del DIF

Los modelos multigrupo requieren algunas restricciones necesarias para la identificación del modelo. Las restricciones para los métodos de calibración independiente se han descrito anteriormente (véase el apartado *Restricciones para la identificación en un modelo multigrupo*). En los métodos de calibración la restricción es dar como conocidos los parámetros de los ítems operativos ($\mathbf{\Delta}_{1(ope)} = \mathbf{\Delta}_{2(ope)}$) y, de este modo, no son necesarias restricciones adicionales.

Ambos procedimientos (tradicionales y de parámetros fijos) se pueden utilizar en el análisis del DIF. Para estudiar el DIF se comparan dos modelos anidados que incluyen las restricciones necesarias para la identificación pero difieren en relación a las restricciones del ítem estudiado.

En el modelo más general no se imponen restricciones adicionales en el ítem pretest bajo estudio ($\mathbf{\Delta}_{1(pre)} \neq \mathbf{\Delta}_{2(pre)}$). En el modelo restringido, se asume que los parámetros del ítem prestest son iguales para ambos grupos ($\mathbf{\Delta}_{1(pre)} = \mathbf{\Delta}_{2(pre)}$). Si el ajuste de ambos modelos difiere significativamente, ello implica presencia de DIF. Por tanto, las restricciones permiten evaluar si la función de respuesta en el ítem pretest es distinta en el grupo de referencia y en el grupo focal.

Las ventajas de los nuevos métodos de DIF, con parámetros fijos para los ítems operativos, son que:

- No requieren reestimar los parámetros de los ítems operativos.
- Los parámetros de los ítems y de distribución se obtienen automáticamente en la métrica original de los ítems operativos.

En relación a los distintos procedimientos de calibración fija se espera que unos procedimientos sean más adecuados que otros. El funcionamiento de un procedimiento resulta más adecuado si:

- Se mantienen las tasas de error tipo I en el valor nominal prefijado por el

investigador.

- Las tasas de potencia (detección del DIF cuando está presente) son mayores.
- La recuperación de los parámetros de los ítems y de la población es más precisa (p.ej., menor RMSE y sesgo).

En concreto, se espera que:

- Si el ítem pretest es un ítem ajustado (ausencia de DIF), los procedimientos más adecuados serán aquellos que hacen uso de la información adicional que proporcionan los ítems pretest (métodos con múltiples ciclos EM).
- Si el ítem pretest es un ítem desajustado (presencia de DIF), los procedimientos más adecuados serán aquellos que no hacen uso de los ítems pretest (métodos con un solo ciclo EM).
- Si la distribución previa inicial es incorrecta (p. ej., presencia de impacto), los procedimientos que no actualizan dicha distribución funcionarán peor (métodos NWU).
- Asumiendo tests de anclaje fiables (algo esperable para el análisis del DIF) y que cada ítem pretest se analiza por separado, las diferencias entre los métodos que utilizan múltiples o un ciclo EM serán pequeñas, considerando que la contribución de un ítem pretest puede tener pocos efectos.

4.8. Anexo

4.8.1. Ignorabilidad de los valores perdidos en un TAI

Nótese que en la ecuación X, se estiman como parámetros aquellos que maximizan $\log f(\mathbf{U}_{(obs)}|\boldsymbol{\Delta},\boldsymbol{\pi})$. Sin embargo, lo correcto es estimar los parámetros que maximizan $\log f(\mathbf{U}_{(obs)},\mathbf{M}|\boldsymbol{\Delta},\boldsymbol{\pi})$. En este anexo se muestra, siguiendo a Mislevy y Wu (1996) que maximizar $f(\mathbf{u}_{i_g(obs)} \mid \theta_{i_g},\boldsymbol{\Delta}_g)$ es equivalente a maximizar $f(\mathbf{u}_{i_g(obs)},\mathbf{m}_{i_g} \mid \theta_{i_g},\boldsymbol{\Delta}_g)$.

En un TAI, la probabilidad conjunta de un patrón de respuestas *completo* y el patrón de valores perdidos puede expresarse como:

$$f(\mathbf{u}_{i_g},\mathbf{m}_{i_g} \mid \theta_{i_g},\boldsymbol{\Delta}_g,\boldsymbol{\Phi}) = f(\mathbf{u}_{i_g} \mid \theta_{i_g},\boldsymbol{\Delta}_g,\boldsymbol{\Phi}) f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g},\theta_{i_g},\boldsymbol{\Delta}_g,\boldsymbol{\Phi}) \tag{4.34}$$

Donde $\boldsymbol{\Phi}$ se refiere a los parámetros del ítem o del evaluado que determinan (probabilísticamente) que una observación se convierta en valor perdido. Puesto que, en TRI, la probabilidad de una respuesta sólo depende de θ_{i_g} y $\boldsymbol{\Delta}_g$:

$$f(\mathbf{u}_{i_g},\mathbf{m}_{i_g} \mid \theta_{i_g},\boldsymbol{\Delta}_g,\boldsymbol{\Phi}) = f(\mathbf{u}_{i_g} \mid \theta_{i_g},\boldsymbol{\Delta}_g) f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g},\theta_{i_g},\boldsymbol{\Delta}_g,\boldsymbol{\Phi}) \tag{4.35}$$

La verosimilitud del patrón de respuestas *observado* y el patrón de valores perdidos puede expresarse como una verosimilitud marginal (i.e., el valor esperado a través de las posibles realizaciones de los valores perdidos):

$$f(\mathbf{u}_{i_g(obs)},\mathbf{m}_{i_g} \mid \theta_{i_g},\boldsymbol{\Delta}_g) = \int f(\mathbf{u}_{i_g(mis)},\mathbf{u}_{i_g(obs)} \mid \theta_{i_g},\boldsymbol{\Delta}_g) f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(mis)},\mathbf{u}_{i_g(obs)},\theta_{i_g},\boldsymbol{\Delta}_g,\boldsymbol{\Phi}) d_{\mathbf{u}_{i_g(mis)}} \tag{4.36}$$

Puesto que en TRI, las respuestas son independientes dados θ_{i_g} y $\boldsymbol{\Delta}_g$:

$$f(\mathbf{u}_{i_g(obs)},\mathbf{m}_{i_g} \mid \theta_{i_g},\boldsymbol{\Delta}_g) = f(\mathbf{u}_{i_g(obs)} \mid \theta_{i_g},\boldsymbol{\Delta}_g) \int f(\mathbf{u}_{i_g(mis)} \mid \theta_{i_g},\boldsymbol{\Delta}_g) f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(mis)},\mathbf{u}_{i_g(obs)},\theta_{i_g},\boldsymbol{\Delta}_g,\boldsymbol{\Phi}) d_{u_{miss}} \tag{4.37}$$

Además se sabe que, si no se utiliza información externa para la selección del primer ítem, la probabilidad de selección de un ítem en un TAI depende sólo de las respuestas observadas a los ítems[4]:

$$f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(mis)}, \mathbf{u}_{i_g(obs)}, \theta_{i_g}, \boldsymbol{\Delta}_g, \boldsymbol{\Phi}) = f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(obs)}) \tag{4.38}$$

Entonces:

$$f(\mathbf{u}_{i_g(obs)}, \mathbf{m}_{i_g} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) = f(\mathbf{u}_{i_g(obs)} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) \int f(\mathbf{u}_{i_g(mis)} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(obs)}) d_{u_{miss}} \tag{4.39}$$

$$f(\mathbf{u}_{i_g(obs)}, \mathbf{m}_{i_g} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) = f(\mathbf{u}_{i_g(obs)} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(obs)}) \int f(\mathbf{u}_{i_g(mis)} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) d_{u_{miss}} \tag{4.40}$$

Y puesto que:

$$\int f(\mathbf{u}_{i_g(mis)} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) d_{u_{miss}} = 1 \tag{4.41}$$

Entonces:

$$f(\mathbf{u}_{i_g(obs)}, \mathbf{m}_{i_g} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) = f(\mathbf{u}_{i_g(obs)} \mid \theta_{i_g}, \boldsymbol{\Delta}_g) f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(obs)}) \tag{4.42}$$

Nótese que $f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(obs)})$ no depende de $\boldsymbol{\Delta}$ ni de θ, por lo que los parámetros

---

[4] En efecto:

$$f(\mathbf{m}_{i_g} \mid \mathbf{u}_{i_g(mis)}, \mathbf{u}_{i_g(obs)}, \theta_{i_g}, \boldsymbol{\Delta}_g, \boldsymbol{\Phi}) = \sum_{s \in T_{i_g}} \prod_{k=1}^{N} \Phi(i_k \mid s_{k-1})$$

donde *k* indica el ítem aplicado en el TAI; s_k indica la secuencia de *k* ítems aplicados y las respuestas a estos ítems; por ejemplo, $s_{30} = ((i_1 = 10, r_1 = 1), (i_2 = 25, r_2 = 0), ..., (i_{30} = 20, r_{30} = 1))$ indicaría que el primer ítem que se aplica es el 10 y es acertado, el segundo es el 25, que es fallado, etc. $\Phi(i_k \mid s_{k-1})$ indica la probabilidad de que un ítem sea aplicado en el paso *k*. En un TAI, la probabilidad de que el ítem i_k sea uno concreto sólo depende del patrón de respuestas a los *k*-1 ítems aplicados. El sumatorio es para todos los patrones de respuesta posibles para los que se cumple que $\mathbf{m}_{i_g}$ es el patrón de ítems respondidos y $\mathbf{u}_{i_g(obs)}$ es el patrón de respuestas observadas. Nótese que $\sum_{s \in T_{i_g}} \prod_{k=1}^{N} \Phi(i_k \mid s_{k-1})$ está definido a priori (según el diseño del TAI) y sólo depende de $\mathbf{u}_{i_g(obs)}$. T_{ig} es el conjunto de secuencias de respuestas con los patrones de patrones de missing y de respuestas al ítem

que maximizan $\log f(\mathbf{u}_{i_g(obs)} \mid \theta_{i_g}, \mathbf{\Delta}_g)$ son los mismos que maximizan $\log f(\mathbf{u}_{i_g(obs)}, \mathbf{m}_{i_g} \mid \theta_{i_g}, \mathbf{\Delta}_g)$.

Capítulo 5

Estudio 1: Detección online del DIF con métodos de Calibración con Parámetros Fijos

5.1. Abstract

En la aplicación a gran escala de tests adaptativos informatizados es común la necesidad periódica de renovar parte del banco de ítems. Para ello, ítems pretest son presentados conjuntamente con ítems operativos. Los ítems pretest son calibrados según algún modelo de TRI y se estudia su posible DIF. En el presente estudio se compara un método de calibración online basado en la imputación de las respuestas faltantes (IM) y tres métodos diferentes de calibración con parámetros fijos, que difieren en número de veces que se actualiza la distribución previa y el número de ciclos EM (OWU-OEM, OWU-MEM y MWU-MEM). Los factores manipulados son: (a) longitud del test; (b) tipo de DIF; (c) tamaño de DIF; (d) tamaño de las muestras; y (e) impacto. En términos generales no se encontraron diferencias entre métodos con respecto al error Tipo I y la potencia en la detección del DIF. Sin embargo, en el método IM se subestimó la varianza del grupo focal, lo que conlleva una sobreestimación del tamaño del DIF. El método MWU-MEM fue el más eficiente en términos de la recuperación de los parámetros y de la distribución de la habilidad de los grupos.

Palabras clave: funcionamiento diferencial del ítem, calibración con parámetros fijos, calibración online, test adaptativos informatizados, test de razón de verosimilitudes de la TRI

Los programas a gran escala de evaluación mediante TAIs requieren de la renovación periódica de partes del banco de ítems (Mills y Stocking, 1996). Para ello, un procedimiento habitual es la calibración online, que consiste en aplicar, a la vez que el test operativo, algunos ítems pretest cuyos parámetros se desconocen. Los examinados no saben qué ítems del test son operativos y cuáles pretest. Sólo los ítems operativos son usados para la estimación del nivel de rasgo (Lei, Chen y Yu, 2006; Segall, 2003; Wainer, 1990).

Una vez conseguida muestra suficiente se procede a la calibración de los ítems pretest y se comprueba que satisfacen ciertas garantías psicométricas, como la ausencia de DIF. Por otro lado, el que la aplicación de los ítems operativos sea adaptativa supone algunas dificultades adicionales en comparación con la calibración de tests fijos: la matriz de datos cuenta con una cantidad importante de valores perdidos y las respuestas a los ítems operativos están basadas en un rango restringido de habilidad (Folk y Golub-Smith, 1996; Haynie y Way, 1995; Hsu, Thompson y Chen, 1998, Stocking, 1988). Adicionalmente, se requiere calibrar los parámetros de los ítems pretest en la métrica de los ítems operativos, de forma que se garantice que las estimaciones del nivel de rasgo que se obtengan utilizando los nuevos ítems sean comparables a las obtenidas con el test operativo. Estas circunstancias plantean la necesidad de revisar los procedimientos convencionales de calibración (Ban, Hanson, Wang, Yi y Harris, 2001), así como los métodos de análisis del DIF (Zwick, 2000, 2007).

5.2. Calibración online en TAIs

En los métodos tradicionales de calibración con parámetros libres (CPL) se calibran conjuntamente todos los ítems, operativos y pretest. La aplicación de CPL resulta problemática por la necesidad de recalibrar los ítems operativos, dada la escasa aplicación de algunos (p.e., Parshall, Kromrey, Harmes y Sentovich, 2001). En los pocos casos en los que se alcanza el criterio de convergencia, CPL tiene un alto coste computacional porque se requiere volver a calibrar cada ítem operativo y las muestras pueden resultar insuficientes, por la elevada frecuencia de valores perdidos (Harmes, Kromrey y Parshall, 2001). Además, existe el riesgo de obtener los parámetros de los ítems pretest en una métrica incorrecta.

Para evitar estos problemas se han propuesto los métodos de Calibración con Parámetros Fijos (CPF). En los procedimientos de CPF se dan como conocidos los parámetros de los ítems operativos. Para ello se fijan los parámetros de los ítems operativos a sus valores previos y se estiman únicamente los parámetros de los ítems pretest. De esta forma se elude el problema de recalibrar los ítems operativos, disminuyendo el tiempo de estimación y eliminándose los problemas de convergencia debidos a la aplicación adaptativa. Además, los ítems pretest se estiman en la escala métrica de los ítems operativos y se requieren muestras de menor tamaño, pues se hace uso de la información previa disponible sobre los parámetros de los ítems operativos (Ban et al., 2001; Ban, Hanson, Yi y Harris, 2002; Kim, 2006).

Para obtener los parámetros de los ítems pretest en la CPF se utiliza la estimación por máxima verosimilitud marginal (MML) usando el algoritmo iterativo EM (Bock y Aitkin, 1981). En cada ciclo *s,* se llevan a cabo dos pasos. Un paso *E* (Esperanza) en el que se estima la distribución posterior de θ y un paso *M* (Maximización) en el que se estiman los parámetros de los ítems, dada la distribución posterior estimada en el paso anterior. En los métodos CPF el elemento central que permite conservar la escala se debe a que los parámetros de los ítems operativos no se estiman en ningún momento. Los parámetros se mantienen fijados a sus valores originales a través del proceso de estimación.

Así, en el paso *E*, la distribución posterior de la habilidad se estima a partir de las respuestas observadas a los ítems ($\mathbf{u}_{i(obs)}$), parte de los cuáles son operativos y tienen parámetros conocidos ($\mathbf{\Delta}_{(ope)}$). En cada ciclo *s* se estima la distribución posterior:

$$f(\theta_k \mid \mathbf{u}_{i(obs)}, \mathbf{\Delta}_{(ope)}, \hat{\mathbf{\Delta}}_{(pre)}^{(s-1)}, \hat{\boldsymbol{\pi}}^{(s-1)}) = \frac{f(\mathbf{u}_{i(obs)} \mid \theta_k, \mathbf{\Delta}_{(ope)}, \hat{\mathbf{\Delta}}_{(pre)}^{(s-1)}) f(\theta_k \mid \hat{\boldsymbol{\pi}}^{(s-1)})}{\sum_{k=1}^{K_g} f(\mathbf{u}_{i(obs)} \mid \theta_{k'}, \mathbf{\Delta}_{(ope)}, \hat{\mathbf{\Delta}}_{(pre)}^{(s-1)}) f(\theta_{k'} \mid \hat{\boldsymbol{\pi}}^{(s-1)})} \tag{5.1}$$

donde $f(\theta_k \mid \hat{\boldsymbol{\pi}}^{(s-1)})$ es *la distribución previa* de la habilidad (definida al inicio de forma arbitraria).

En el paso *M*, se estiman sólo los parámetros de los ítems pretest, que maximizan la función de log-verosimilitud:

$$\mathbf{\Delta}_{(pre)}^{(s)} = \arg\max_{\mathbf{\Delta}} \phi(\mathbf{\Delta}_{(pre)}) =$$
$$\arg\max_{\mathbf{\Delta}} \left\{ \sum_{i=1}^{N} \sum_{k=1}^{K} \left\{ \log\left[f(\mathbf{u}_{i(obs)(pre)} \middle| \mathbf{\Delta}_{g(pre)}, \theta_k) \right] f(\theta_k \mid \mathbf{u}_{i(obs)}, \mathbf{\Delta}_{(ope)}, \hat{\mathbf{\Delta}}_{(pre)}^{(s-1)}, \hat{\boldsymbol{\pi}}^{(s-1)}) \right\} \right\} \tag{5.2}$$

Y se actualiza la distribución previa:

$$\boldsymbol{\pi}^{(s)} = \arg\max_{\boldsymbol{\pi}} \psi(\boldsymbol{\pi}) =$$
$$\arg\max_{\boldsymbol{\pi}} \left\{ \sum_{i=1}^{N} \sum_{k=1}^{K} \left\{ \log\left[f(\theta_k | \boldsymbol{\pi}) \right] f(\theta_k \mid \mathbf{u}_{i(obs)}, \boldsymbol{\Delta}_{(ope)}, \hat{\boldsymbol{\Delta}}_{(pre)}^{(s-1)}, \hat{\boldsymbol{\pi}}^{(s-1)}) \right\} \right\} \tag{5.3}$$

El algoritmo anterior expresa el caso más general (método MWU-MEM). La variación en número de veces que se actualiza la distribución previa y el número de ciclos EM utilizados para la estimación permiten definir diversas aproximaciones a los métodos de CPF.

Específicamente, respecto a la distribución previa es posible:

1) No actualizarla (NWU), asumiendo una distribución previa inicial, por ejemplo:

$$\boldsymbol{\pi}^{(s)} = \boldsymbol{\pi}^{(0)} \tag{5.4}$$

donde se asume que la distribución previa de θ_k, $f(\theta_k \mid \hat{\boldsymbol{\pi}}^{(0)})$, es la distribución normal estándar, $N(0,1)$; es decir, $\boldsymbol{\pi}^{(0)} = \{0,1\}$.

2) Actualizarla una única vez utilizando únicamente la información de los ítems operativos (OWU). En ese caso

$$\boldsymbol{\pi}^{(1)} = \arg\max_{\boldsymbol{\pi}} \psi(\boldsymbol{\pi}) =$$
$$\arg\max_{\boldsymbol{\pi}} \psi(\boldsymbol{\pi}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \left\{ \log\left[f(\theta_k | \boldsymbol{\pi}) \right] f(\theta_k \mid \mathbf{u}_{i(obs)(ope)}, \boldsymbol{\Delta}_{(ope)}, \hat{\boldsymbol{\pi}}^{(0)}) \right\} \tag{5.5}$$

3) Actualizarla múltiples veces utilizando la información tanto de los ítems operativos como de los ítems pretest (MWU).

Por otro lado, en relación con el número de ciclos EM, los métodos pueden emplear:

1) Un ciclo EM (OEM), de forma que en el paso *E* se calcula la distribución posterior a partir de los ítems operativos:

$$f(\theta_k \mid \mathbf{u}_{i(obs)}, \boldsymbol{\Delta}_{(ope)}, \hat{\boldsymbol{\pi}}^{(s-1)}) = \frac{f(\mathbf{u}_{i(obs)(ope)} \mid \theta_k, \boldsymbol{\Delta}_{(ope)}) f(\theta_k \mid \hat{\boldsymbol{\pi}}^{(s-1)})}{\sum_{k=1}^{K} f(\mathbf{u}_{i(obs)(ope)} \mid \theta_{k'}, \boldsymbol{\Delta}_{(ope)}) f(\theta_{k'} \mid \hat{\boldsymbol{\pi}}^{(s-1)})} \tag{5.6}$$

2) Múltiples ciclos EM (MEM), de forma que, a partir del segundo paso *E,* se calcula la distribución posterior utilizando las respuestas a todos los ítems, incluidos los ítems pretest.

5.3. Detección del DIF en TAIs

El uso de técnicas de DIF en el contexto de los TAIs se enfrenta a las mismas dificultades descritas para los procedimientos de calibración. La presencia de valores perdidos, así como la necesidad de obtener medidas de tamaño del efecto y del impacto en la métrica del TAI original, hacen necesaria la adaptación de estas técnicas.

El uso de un procedimiento basado en la TRI resulta la opción más natural en este contexto, pues esta se aplica a lo largo de todo el proceso de elaboración del TAI (p.e., la calibración del banco de ítems, la selección de los ítems mediante el algoritmo adaptativo o la puntuación de las respuestas). En el presente trabajo se analiza el funcionamiento del test de razón de verosimilitudes de la TRI (*IRT LRT*). En este procedimiento, se asume un modelo de TRI y se contrasta la equivalencia en los parámetros del ítem estudiado a través de los grupos. Para ello se compara estadísticamente el ajuste de dos modelos multigrupo: a) un modelo en el que todos los parámetros son equivalentes a través de los grupos; b) un modelo en el que los parámetros del ítem estudiado *pueden* diferir a través de los grupos.

El IRT LRT satisface los criterios de eficiencia descritos por Wainer, Bradlow y Wang (2010): 1) *precisión* en la identificación de la presencia, naturaleza y tamaño del DIF; 2) *capacidad* de usar la información disponible que permita trabajar con muestras lo más pequeñas posibles; 3) *funcionalidad* en la implementación del análisis; 4) *flexibilidad* en la detección en situaciones novedosas; y 5) *robustez* a violaciones de los supuestos de los análisis. Así, IRT LRT es uno de los procedimientos más eficientes en el control de la tasa de error Tipo I (Cohen, Kim y Wollack, 1996; Kim y Cohen, 1998; Lei et al., 2006; Lopez-Rivas, Stark y Chernyshenko, 2008; Wang y Yeh, 2003), puede ser aplicado tanto para la detección del DIF unidireccional como para la detección del DIF no-unidireccional y la funcionalidad de su uso se ha incrementado gracias a la disponibilidad de programas como el IRTLRTDIF v.2.0 (Thissen, 2001).

A pesar de lo anterior la aplicación del IRT LRT a la detección del DIF en TAIs se ha visto complicada por la dificultad de tratar con matrices dispersas de datos. Lei

et.al. (2006) propusieron resolver el problema de los valores perdidos en el TAI imputando las respuestas faltantes (las probabilidades de acierto a esos ítems se imputaban asumiendo como verdadero el nivel de rasgo estimado en el TAI). En términos generales, IRT-LRT proporcionaba un adecuado control de la tasa de error tipo I y tasas de potencia que dependían del tamaño del DIF y las características de los ítems pretest (tasas más altas con DIF grande e ítems muy discriminativos y fáciles).

Aunque la aproximación de Lei et al. (2006) es interesante, su propuesta resulta algo ineficiente (aumenta el coste computacional por la estimación de los parámetros en los ítems operativos) y la implementación en el programa IRTLRDIF no es óptima (p.e., el número de ítems operativos para el anclaje está limitado a 200 y, en su estudio, fue necesario reducir el número de ítems operativos a 84). Por otro lado, los autores no proporcionan medidas de tamaño del efecto ni del impacto estimado, existiendo el riesgo de que estas se recuperen en una métrica inadecuada. Finalmente, un problema adicional es que en el proceso de imputación se asume el nivel de rasgo estimado como verdadero; dado el ingente número de respuestas imputadas esto puede sesgar los resultados obtenidos, especialmente en aquellos casos en los que la estimación del rasgo sea poco precisa o esté sesgada.

Todos los problemas anteriores tienen que ver con que, al aplicar el IRT LR, se utiliza la CPL, lo que requiere recalibrar los ítems operativos. En el presente estudio se propone una estrategia alternativa: aplicar el IRT LR dando como fijos los parámetros de los ítems del TAI operativo; es decir, utilizar un método CPF.

5.4. Detección online del DIF en TAIS con métodos de CPF

La extensión multigrupo de los procedimientos de CPF permite aplicar la prueba IRT LR para la detección online del DIF de ítems pretest, sin necesidad recalibrar los ítems del TAI. En la actualidad el único programa en el que se pueden implementar los diferentes métodos de CPF en un estudio multigrupo es ICL (Hanson, 2002), de libre distribución. En este caso, se trataría de comparar estadísticamente el ajuste de dos modelos multigrupo, un modelo *a* en el que los parámetros del ítem estudiado son equivalentes a través de los grupos y un modelo *b* en el que los parámetros del ítem estudiado *pueden* diferir a través de los grupos. Existen dos diferencias a considerar con respecto al método tradicional:

1) en el procedimiento usual los parámetros de los ítems operativos se estiman. En el procedimiento propuesto los parámetros de los ítems operativos se fijan.
2) en el procedimiento usual se fija la media y la desviación típica del grupo de referencia (a 0 y 1) y se estiman los parámetros de distribución en el grupo focal. En el procedimiento propuesto, los parámetros de distribución de ambos grupos se estiman.

En relación a qué método de CPF aplicar, el estudio de Kim (2006) ofrece algunas direcciones en el contexto de la calibración online. En primer lugar, el uso de métodos que no actualizan la distribución de la habilidad –p.e., N(0,1)– es desaconsejable ya que ofrece resultados incorrectos si la distribución de la nueva muestra es otra. En relación al resto de los métodos, Kim (2006) sugiere que actualizar la distribución de la habilidad (ya sea una o múltiples veces) y, sobre todo, emplear múltiples ciclos EM, resulta más eficiente, especialmente si los ítems operativos se han calibrado en muestras pequeñas. Esto es porque en tales situaciones los ítems pretest contribuyen a actualizar la distribución previa y la distribución posterior. Esta última conclusión puede no ser generalizable al contexto de detección online del DIF ya que el ítem *pretest* es un candidato a ítem *desajustado.* En ese caso, el uso del ítem pretest para actualizar las distribuciones previa y posterior puede resultar incorrecto si no se toman las precauciones adecuadas. Nuestra propuesta es resolver este problema estimando las distribuciones de los grupos en el modelo *b* (que "siempre" es correcto) y asumiendo esas mismas distribuciones para el modelo *a*. De esta forma se garantiza además que los modelos están anidados (Woods, 2008a). En cualquier caso, el efecto de la calibración de un único ítem pretest puede hacer que las diferencias entre los métodos sean pequeñas.

En el presente estudio se comparan tres métodos CPF (OWU-OEM, OWU-MEM y MWU-MEM) para detectar el DIF de un ítem pretest. Al tratarse sólo de un ítem pretest se esperan pocas diferencias entre los métodos, especialmente en las condiciones en las que no haya impacto y la muestra sea pequeña. Atendiendo a los resultados de Kim (2006), es esperable que funcionen algo mejor los métodos que hacen uso del ítem pretest para actualizar la distribución posterior (OWU-MEM) o, mejor aún, la posterior y la previa (MWU-MEM).

Finalmente, se compararan los métodos de CPF en los que se actualiza la distribución previa con la propuesta de Lei et al. (2006) de imputar las respuestas (IM). Puesto que la propuesta de Lei et al. (2006) de aplicar CPL con imputación de respuestas es poco eficiente en tiempo de computación, el presente estudio hemos modificado su implementación, combinando la imputación de respuestas con el mejor método de calibración fija (MWU-MEM). De esta forma, podemos aislar mejor los efectos de la imputación de respuestas en la detección del DIF. Se espera que el método IM proporcione menor potencia, peor recuperación de las distribuciones del nivel de rasgo y de los tamaños del efecto, especialmente en TAIs de menor longitud en los que la estimación de la habilidad se realiza con mayor error.

5.5. Método

5.5.1. Condiciones de aplicación del TAI

Para la simulación de aplicaciones adaptativas de una prueba se empleó el programa CATSIM 1.0 (Cuevas, Abad, Olea, Barrada y Garrido, 2010). A cada examinado se le administraba un TAI de longitud fija con ítems seleccionados de entre los 300 ítems operativos. El nivel de rasgo inicial era asignado al azar dentro del intervalo (–0.4, 0.4). Se empleó estimación MAP con una distribución normal estándar como distribución a priori. Como regla de selección de ítems se empleó el método progresivo (Revuelta y Ponsoda, 1998) según el cual la selección de ítems tiene un elevado componente aleatorio al comienzo del test y la importancia de la información de Fisher en la selección se va incrementando según avanza el test. Tras el TAI, todos los examinados recibieron los mismos 23 ítems pretest.

5.5.2. Parámetros de los ítems operativos

Los parámetros de los ítems operativos eran iguales para el grupo de referencia y el grupo focal. Las distribuciones de los parámetros se aproximan a las de los ítems del Law School Admission Test y se generaron siguiendo la información proporcionada por Nandakumar y Roussos (2004). Para ítems con parámetro *b* menor o igual a –1, el parámetro *a* seguía una distribución lognormal(–0.357, 0.25), acotados dentro del rango

[0.4, 1.1]. Para el resto de ítems, la distribución del parámetro *a* es lognormal(–0.223, 0.34) dentro del rango [0.4, 1.7]. El parámetro *b* seguía una distribución *N*(0, 1) dentro del rango [–3, 3]. El parámetro *c* seguía una distribución uniforme entre [0.12, 0.22].

5.5.3. Parámetros de los ítems pretest

El número de ítems pretest fue de 23. De estos, 10 tenían DIF unidireccional, 6 tenían DIF no-unidireccional y 7 no tenían DIF. En las condiciones con DIF, la mitad de los ítems presentaban DIF moderado (β = 0.05) y la otra mitad tenía DIF grande (β = 0.10). β es una medida de tamaño del efecto del DIF (Wainer, 1990) calculada como:

$$\beta = \int |P(u_{iR} = 1 \mid \theta) - P(u_{iF} = 1 \mid \theta)| g(\theta) \tag{5.7}$$

donde $P(u_{iR} = 1 \mid \theta)$ y $P(u_{iF} = 1 \mid \theta)$ son las probabilidades de acierto del ítem para los grupos de referencia y focal y $g(\theta)$ es la densidad de θ asumiendo una distribución normal típica. En otras palabras, β es el valor esperado de la diferencia en la probabilidad de acierto entre grupos.

Tabla 5.1. Parámetros para generar los ítems en los que se estudia el DIF

	SIN DIF			DIF UNIDIRECCIONAL					DIF NO UNIDIRECCIONAL			
Item	$a_R=a_F$	$b_R=b_F$	Item	$a_F=a_R$	b_F	b_R	β	Item	a_F	a_R	$b_R=b_F$	β
1	0.70	-1.30	8	0.70	-1.30	-1.670	0.05	18	0.80	1.22	-1.30	0.05
2	0.70	1.30	9	0.70	1.30	0.955	0.05	19	0.80	1.24	1.30	0.05
3	0.80	0.00	10	1.20	-1.30	-1.650	0.05	20	0.80	1.19	0.00	0.05
4	1.20	-1.30	11	1.20	1.30	0.985	0.05	21	0.80	2.23	-1.30	0.10
5	1.20	1.30	12	0.80	0.00	-0.240	0.05	22	0.80	2.65	1.30	0.05
6	0.80	-1.30	13	0.70	-1.30	-2.130	0.10	23	0.80	1.83	0.00	0.10
7	0.80	1.30	14	0.70	1.30	0.655	0.10					
			15	1.20	-1.30	-2.170	0.10					
			16	1.20	1.30	0.725	0.10					
			17	0.80	0.00	-0.485	0.10					

- Ítems con DIF unidireccional:

  En primer lugar, se generaron los ítems para el grupo focal. Posteriormente, se cambio el parámetro *b* del grupo de referencia ajustándolo para obtener el tamaño de DIF deseado. De los 10 ítems, 8 se obtienen de la combinación de valor en el parámetro *a* (0.7 ó 1.2), valor en el parámetro b (–1.3 ó 1.3) y tamaño del DIF. Los otros dos ítems representan un ítem promedio del LSAT (*a* = 0.8; *b* = 0) con dos tamaños de DIF diferente. Los ítems para el grupo de referencia eran más fáciles que para el grupo focal, puesto que su parámetro *b* era menor.
- Ítems con DIF no-unidireccional:

  Los ítems para el grupo de referencia se generaron con un parámetro a constante igual a 0.8 y con tres niveles de parámetro b (–1.3, 0, 1.3). Al cruzar esto con los dos tamaños de DIF tenemos los 6 ítems de esta condición. El parámetro a para el grupo focal se aumentó hasta conseguir los dos tamaños de DIF deseados.
- Ítems sin DIF:

  Los 7 ítems son los diferentes ítems empleados con el grupo de referencia en las condiciones anteriores.

Para todos los ítems pretest, el parámetro *c* se fijó en 0.17. Los valores de los parámetros de los ítems estudiados se consideran dentro del rango observado en test reales (Lopez-Rivas, et al., 2008).

5.5.4. Factores manipulados

Se manipularon los siguientes factores:

- Longitud del TAI (10 ó 30 ítems).
- Impacto: Se generaron dos condiciones, sin impacto y con impacto. En la primera el nivel de habilidad seguía una distribución normal estándar en ambos grupos. En la segunda, la distribución del grupo focal tenía media menor, N(–0.5, 1). Estas distribuciones son comunes en contextos aplicados y en estudios de simulación (Finch y French, 2007,French y Finch, 2008; Lei, et al., 2006; Stark, Chernyshenko y Drasgow, 2006).

- Tamaño de las muestras: El grupo de referencia y focal tienen el mismo tamaño muestral. Se incluyeron dos condiciones, con 250 ó 500 examinados por grupo. Emplear 500 evaluados por grupo se considera un tamaño recomendable para el análisis del DIF en test fijos en procedimientos basados en la TRI (Clauser y Mazor, 1998). Las muestras pequeñas son cercanas a las utilizadas por Kim (2006) como un ejemplo de situación realista en el contexto de la calibración online.

Cada una de estas ocho condiciones diferentes (2 × 2 × 2) fue replicada 100 veces. En cada una de estas réplicas se administraron los 23 ítems pretest que variaban según (ver Tabla 5.1):

- Tipo de DIF: Sin DIF, DIF unidireccional y DIF no unidireccional.
- Tamaño del DIF: Pequeño ($\beta = 0.05$) y grande ($\beta = 0.10$).

5.5.5. Métodos de calibración

Cada uno de los 23 ítems pretest fue calibrado separadamente según cuatro procedimientos, utilizando conjuntamente las respuestas a los ítems operativos del TAI y al ítem pretest. En todos los casos, los parámetros de los ítems operativos se fijaban a sus valores reales (calibración con parámetros fijos). Para cada examinado teníamos dos vectores de respuestas, el empírico (con valores faltantes para los ítems no administrados) y el imputado (donde se imputaban las respuestas a los ítems no administrados en el TAI). Para la imputación la probabilidad de acierto a los ítems era calculada mediante los parámetros de los ítems y el nivel de rasgo estimado al finalizar el TAI. Para el caso de vectores sin datos imputados, se pusieron a prueba tres métodos diferentes: a) una sola actualización de la distribución previa y un ciclo EM (OWU-OEM); b) una sola actualización de la distribución previa y múltiples ciclos EM (OWU-MEM); y c) múltiples actualizaciones de la distribución previa y múltiples ciclos EM (MWU-MEM). Para los vectores con respuestas imputadas se usó el método MWU-MEM, evaluando así la propuesta de Lei, et al. (2006), con la diferencia de usar CPF en lugar de CPL.

El software de calibración empleado fue el ICL (Hanson, 2002). Se definieron distribuciones previas para los parámetros *a* y *c*: para el *a* lognormal(0, 0.5); para el *c* beta(4, 16, 0, 1) donde los primeros dos valores son los parámetros para la forma y los dos siguientes el límite inferior y superior. Estas distribuciones son comunes a otros programas de estimación como PARSCALE o BILOG (du Toit, 2003). Se estableció un máximo de 1000 iteraciones EM para los procedimientos con MEM y un criterio de convergencia de 0.001.

5.5.6. Detección del DIF

Para analizar el DIF en cada ítem se calculó el logaritmo de verosimilitud de dos modelos: 1) un modelo en el que se restringen los parámetros de todos los ítems a ser iguales entre el grupo de referencia y el grupo focal (con función de log-verosimilitud LL_{igual}); y 2) un modelo en el que se mantienen los parámetros de los ítems operativos iguales y se permite que los parámetros del ítem pretest varíe entre los grupos (con función de log-verosimilitud $LL_{diferente}$). Se calculó el estadístico de prueba G^2 como $G^2 = -2(LL_{igual}) - (-2LL_{diferente})$ que se distribuye según χ^2 con 3 grados de libertad (la diferencia en el número de parámetros a estimar). Una prueba estadísticamente significativa indica la presencia de DIF. Se obtuvieron adicionalmente los parámetros de los ítems y de la distribución para cada grupo a partir del modelo libre ($LL_{diferente}$).

5.5.7. Criterios de valoración

La efectividad relativa de los cuatro métodos de calibración (tres para vectores sin imputación, uno para respuestas faltantes imputadas) se evaluó según los siguientes criterios:

- Detección del DIF:

  Para evaluar la efectividad en la detección del DIF se obtuvo la potencia y el error Tipo I. Debido a que la potencia puede aumentar de manera espuria cuando las tasas de error Tipo I son muy altas, se estableció un valor crítico para las tasas de error Tipo I de 0.05 con un IC95% [0.007, 0.093] (i.e.,

$0.05 \pm 1.96\sqrt{0.05(1-0.05)/100}$). Por otro lado, consideraremos una potencia como aceptable si es igual o mayor a 0.80.

- Recuperación de los parámetros de distribución en cada grupo:
  Se calculó la media y desviación típica de la distribución posterior para las diferentes muestras de los grupos de referencia y focal. La calidad de la recuperación se evaluó mediante el *sesgo*.
- Recuperación de los parámetros de los ítems en cada grupo:
  Se calculó el sesgo y el RMSE de la recuperación de cada uno los parámetros.
- Recuperación del tamaño del DIF (β).

La calidad de la recuperación de β se evaluó mediante el sesgo y el RMSE.

5.5.8. Análisis

Se realizaron ANOVAs para determinar qué factores afectaban a la tasa de error Tipo I, a la potencia y a la recuperación de β. Se utilizó un ANOVA mixto en el que los procedimientos de estimación empleados se consideran una medida repetida, mientras que las variables longitud del TAI, tamaño de la muestra, impacto y tamaño del DIF se consideran medidas independientes. Estos análisis se realizaron por separado para el DIF unidireccional y no-unidireccional. La variable dependiente era la proporción de veces que se detectaba DIF a través de las réplicas (para el análisis de las tasas de error tipo I y de potencia) o el sesgo promedio a través de las réplicas en la recuperación de β. En el modelo de ANOVA, sólo se consideraron las interacciones dobles entre los factores inter-sujeto.

Todos los efectos informados son significativos con $p < 0.05$. Para evaluar el tamaño del efecto se utilizó como medida η^2 parcial. Únicamente se informa de aquellos efectos con tamaños del efecto superiores a .14 que, en la clasificación de Cohen (1988, 1992), son considerados grandes.

Adicionalmente, se llevo a cabo un ANOVA para estudiar si el tipo de ítem (sin DIF, DIF unidireccional pequeño, DIF unidireccional grande, DIF no unidireccional pequeño y DIF no unidireccional grande) tenía efectos significativos en la recuperación de los parámetros de distribución, considerando como factores adicionales tamaño de las muestras, longitud del TAI, impacto y método.

5.6. Resultados

5.6.1. Detección del DIF

Tasas de error Tipo I

La Tabla 5.2 contiene la proporción de falsos positivos (tasas de error tipo I) para cada método en cada una de las condiciones. En general las tasas de error estuvieron dentro del rango establecido. El ANOVA mostró el efecto significativo del método de estimación [$F(3,147)=13.60$, $\eta_p^2 = .22$], obteniéndose mayores tasas de error tipo I para el método IM ($\bar{x}_{O-O} = .021; \bar{x}_{O-M} = .023; \bar{x}_{M-M} = .023; \bar{x}_{IM} = .028$).

Tabla 5.2. Tasas de error Tipo I para todas las combinaciones de las condiciones en el estudio

		Distribución N(0,1) Métodos				N(-0.5,1) Métodos			
Tamaño de la muestra	Longitud del test	O-O	O-M	M-M	IM	O-O	O-M	M-M	IM
250-250									
	10	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.02
	30	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.02
500-500									
	10	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.04
	30	0.02	0.02	0.02	0.03	0.02	0.02	0.03	0.03

Nota: En los métodos la primera letra representa el número de veces que se actualiza la distribución (**O** = una vez y **M** = múltiples veces), la segunda letra representa el número de ciclos EM (**O** =un ciclo y **M**= múltiples ciclos). **IM** = Imputación.

Potencia: DIF unidireccional

En la parte superior de la Tabla 5.3 se presentan las tasas de potencia promedio para detectar el DIF unidireccional.

Tabla 5.3. Tasas de Potencia para el DIF unidireccional y no-unidireccional

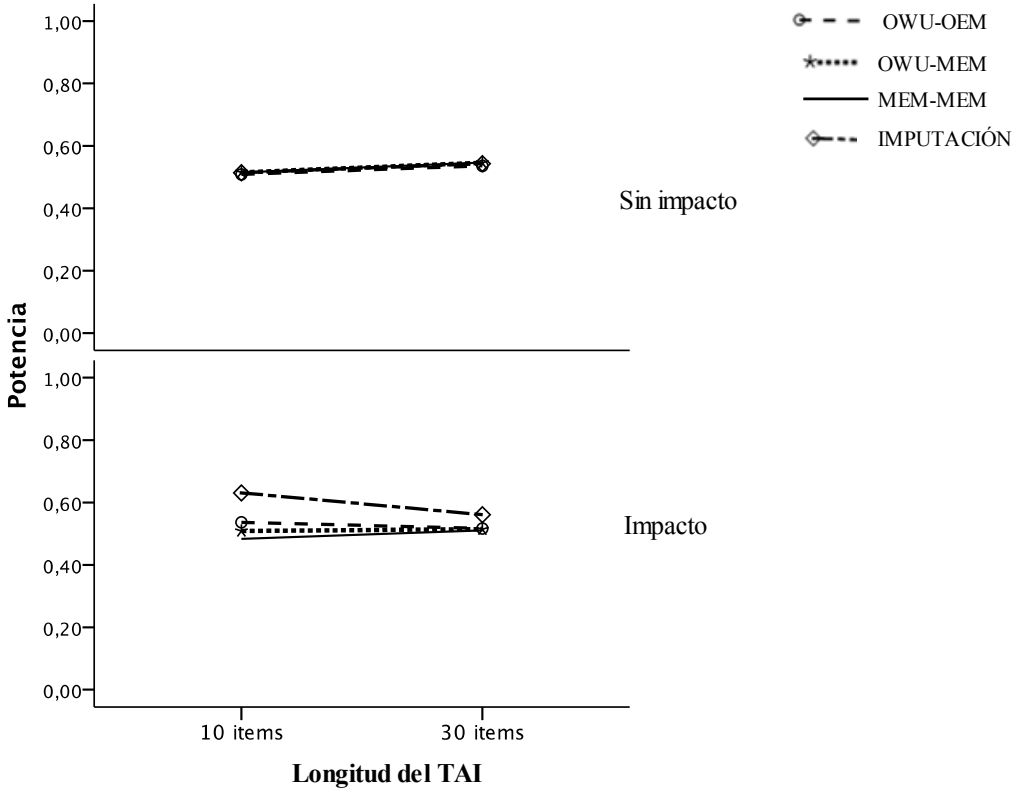
β	Tamaño de las muestras	Longitud del TAI	Impacto				Sin impacto			
			O-O	O-M	M-M	IM	O-O	O-M	M-M	IM
			DIF Unidireccional							
0.05										
	250-250	10	0.14	0.14	0.15	0.16	0.20	0.17	0.15	0.30
		30	0.15	0.16	0.16	0.16	0.18	0.18	0.18	0.21
	500-500	10	0.32	0.33	0.33	0.32	0.40	0.36	0.31	0.56
		30	0.39	0.40	0.40	0.39	0.34	0.34	0.34	0.42
0.10										
	250-250	10	0.67	0.67	0.67	0.68	0.65	0.64	0.62	0.73
		30	0.69	0.69	0.69	0.69	0.64	0.64	0.64	0.68
	500-500	10	**0.91**	**0.91**	**0.91**	**0.90**	**0.89**	**0.87**	**0.85**	**0.94**
		30	**0.93**	**0.93**	**0.93**	**0.93**	**0.90**	**0.89**	**0.89**	**0.92**
			DIF No-unidireccional							
0.05										
	250-250	10	0.12	0.14	0.14	0.14	0.15	0.16	0.16	0.21
		30	0.19	0.20	0.20	0.20	0.17	0.18	0.18	0.20
	500-500	10	0.31	0.37	0.37	0.36	0.31	0.33	0.33	0.41
		30	0.38	0.41	0.42	0.42	0.35	0.35	0.35	0.36
0.10										
	250-250	10	0.56	0.64	0.65	0.64	0.52	0.56	0.57	0.61
		30	0.68	0.71	0.71	0.72	0.67	0.71	0.71	0.72
	500-500	10	**0.86**	**0.90**	**0.89**	**0.90**	**0.83**	**0.87**	**0.88**	**0.88**
		30	**0.95**	**0.97**	**0.97**	**0.96**	**0.92**	**0.94**	**0.94**	**0.92**

Nota: **O-O** = OWU-OEM, **O-M**=OWU-MEM, **M-M**=MWU-MEM e **IM** = Imputación

La potencia se incrementa con el tamaño de DIF [$F(1,69)=342.17$, $\eta_p^2 = .83$] y el tamaño de las muestras [$F(1,69)=60.89$, $\eta_p^2 = .47$]; la potencia era adecuada cuando el tamaño del DIF y el tamaño muestral eran grandes. En relación a los métodos, se

encontró un efecto del método de estimación [$F(3,207) = 67.16$; $\eta_p^2 = .49$] y de la interacción de este con la presencia de impacto [$F(3,207) = 71.35$; $\eta_p^2 = .51$], con la longitud del TAI [$F(3,207) = 16.65$; $\eta_p^2 = .19$] y con ambas variables [$F(3,207) = 14.69$; $\eta_p^2 = .18$]. En la Figura 5.1 se representa la potencia como función del impacto, la longitud y el método de estimación. En ausencia de impacto no hay diferencias entre los métodos ($\bar{x}_{O-O} = .52; \bar{x}_{O-M} = .53; \bar{x}_{M-M} = .53; \bar{x}_{IM} = .53$). Sin embargo y de manera inesperada se encontró que en presencia de impacto la potencia al utilizar el método IM es superior a la obtenida con los otros métodos, especialmente si el test es corto (en presencia de impacto, si el TAI es corto: $\bar{x}_{O-O} = .54; \bar{x}_{O-M} = .51; \bar{x}_{M-M} = .48; \bar{x}_{IM} = .63$).

Figura 5.1. Potencia para cada método, en función del impacto y la longitud del TAI



Potencia: DIF no-unidireccional

La potencia para el DIF no-unidireccional se muestra en la parte inferior de la Tabla 5.3. El patrón de resultados es similar al descrito para el DIF unidireccional: la potencia se incrementa con el tamaño de DIF [$F(1,37) = 138.48$, $\eta_p^2 = .79$] y el tamaño de las muestras [$F(1,37) = 27.31$, $\eta_p^2 = 0.43$]. También se encontró un efecto del método [$F(3,37) = 21.68$; $\eta_p^2 = .37$] pero no del resto de los factores (longitud del TAI e impacto). El método O-O produce menor potencia que el resto de los métodos [$\bar{x}_{O-O} = .50$; $\bar{x}_{O-M} = .52$; $\bar{x}_{M-M} = .53$; $\bar{x}_{IM} = .54$].

5.6.2. Efectos del método en la recuperación de la escala

En primer lugar, cabe señalar que no hubo efectos estadísticamente significativos en la recuperación de las medias del tipo de ítem pretest ($p > .05$) o de sus interacciones con el resto de los factores. En relación a las desviaciones típicas sólo se encontró un efecto pequeño para la interacción entre método y tipo de ítem al estimar la desviación típica del grupo de referencia. Sin embargo, el efecto era muy pequeño desde el punto de vista práctico. Por ello, para simplificar los resultados, la Tabla 5.4 presenta los resultados de la recuperación de la media y la desviación típica de los grupos de Referencia y Focal para cada uno de los métodos únicamente en función del tamaño de las muestras, la longitud del TAI y el impacto. En la parte superior de la tabla se muestra la condición sin impacto N(0,1) para ambos grupos, y en la parte inferior la distribución del grupo focal es de N(-0.5,1), la del grupo de referencia es la misma que en la parte superior.

Tabla 5.4. Medias y desviaciones típicas promedio de la distribución de la habilidad para el grupo de referencia y el grupo focal

Tamaño	Longitud	Media				Desviación Típica			
muestras	del TAI	O-O	O-M	M-M	IM	O-O	O-M	M-M	IM
				Sin impacto					
		Grupo de Referencia							
250-250	10	0.00	0.00	0.00	0.04	1.00	1.00	1.00	0.88
	30	0.00	0.00	0.00	0.01	1.00	1.00	1.00	0.97
500-500	10	0.00	0.00	0.00	0.04	1.00	1.00	1.00	0.88
	30	0.00	0.00	0.00	0.01	1.00	1.00	1.00	0.97
		Grupo Focal							
250-250	10	0.00	0.00	0.00	0.03	1.00	1.00	1.00	0.88
	30	0.00	0.00	0.00	0.01	1.00	1.00	1.00	0.97
500-500	10	0.01	0.01	0.01	0.04	1.00	1.00	1.00	0.89
	30	0.01	0.01	0.01	0.02	1.00	1.00	1.00	0.97
				Impacto					
		Grupo de Referencia							
250-250	10	-0.01	-0.01	-0.01	0.03	0.99	0.99	0.99	0.88
	30	0.00	0.00	0.00	0.02	1.00	1.00	1.00	0.97
500-500	10	0.00	0.00	0.00	0.04	1.00	1.00	1.00	0.89
	30	0.01	0.01	0.01	0.02	1.00	1.00	1.00	0.97
		Grupo Focal							
250-250	10	-0.40	-0.40	-0.51	-0.36	0.97	0.97	0.99	0.86
	30	-0.46	-0.46	-0.51	-0.46	0.98	0.98	0.99	0.96
500-500	10	-0.40	-0.40	-0.50	-0.35	0.98	0.98	1.01	0.87
	30	-0.46	-0.46	-0.50	-0.45	0.98	0.98	0.99	0.95

O-O = OWU-OEM, **O-M**=OWU-MEM, **M-M**=MWU-MEM e **IM** = Imputación.

Todos los efectos relacionados con el método y con las interacciones de este y los otros factores (tamaño muestral, longitud del TAI e impacto) eran estadísticamente significativos y, la mayor parte, grandes ($\eta_p^2 > .14$). Describiremos pues los efectos más importantes desde el punto de vista práctico. Se encuentra que al utilizar la imputación

(IM) se subestima la variabilidad en el nivel de rasgo en ambos grupos y, en presencia de impacto, se subestima el impacto, ya que se sobreestima la media del grupo focal en el nivel de rasgo. Estos sesgos, que dependen poco del tamaño muestral, son mayores cuanto menor la longitud del TAI.

Con respecto a la comparación de los otros métodos de calibración fija, los resultados dependen de la presencia de impacto y de la longitud del TAI. En ausencia de impacto no se encuentran diferencias relevantes en la recuperación de las distribuciones de los grupos. Sin embargo, en presencia de impacto, M-M produce una mejor recuperación de la media en el grupo focal que O-O y O-M. Estas diferencias se reducen al aumentar la longitud del TAI.

5.6.3. Recuperación de los parámetros de los ítems pretest

En las Figuras 5.3, 5.4 y 5.5 se presentan el sesgo y el RMSE para la recuperación de los parámetros de los ítems pretest en el grupo de Referencia. Los resultados muestran que en el grupo de referencia el parámetro *a* se subestima y el parámetro *b* se sobreestima independientemente de la longitud del test, del tamaño muestral y de la presencia de impacto (ver figura 5.2 y 5.3). El método que peor funciona es el método O-O sobre todo cuando el TAI es corto. Esto se refleja también en el RMSE.

En las Figuras 5.6, 5.7 y 5.8 se presenta el sesgo y el RMSE para la recuperación de los parámetros *a*, *b* y *c* en el grupo Focal. En relación a los parámetros del grupo focal, se replica el peor funcionamiento del método O-O (subestimación del parámetro *a* y sobreestimación del parámetro *b*, especialmente cuando la longitud del TAI es pequeña) aunque, en este caso, el método IM produce el mismo patrón de resultados en la estimación del parámetro *b*. Para ambos métodos (O-O y IM), la sobreestimación de *b* era mayor en las condiciones de impacto, especialmente si el TAI era corto.

En general los métodos M-M y O-M obtuvieron medidas de sesgo y de RMSE menores respecto a O-O y IM.

5.6.4. Recuperación del Tamaño del DIF

Ausencia de DIF (β = 0)

Los resultados del sesgo y RMSE en la estimación de β del ítem pretest en ausencia de DIF se presentan en la tabla 5.5. Se encontró que el sesgo se reducía aumentar el tamaño muestral [$F(1,49) = 27.51$; $\eta_p^2 = .36$; $\mu_{250-250} = .039$; $\mu_{500-500} = 0.030$]; el efecto de método era significativo [$F(3,147) = 70.33$; $\eta_p^2 = .46$] y también la interacción con el impacto [$F(3,147) = 18.38$; $\eta_p^2 = .27$]. En presencia de impacto, apenas hay diferencias entre los métodos ($\mu_{O-O} = \mu_{O-M} = \mu_{M-M} = .035$; $\mu_{IM} = .036$). En ausencia de impacto, el sesgo de O-O es menor ($\mu_{O-O} = .031$; $\mu_{O-M} = \mu_{M-M} = \mu_{IM} = .034$). En cualquier caso, las diferencias entre los métodos son de poca relevancia práctica. En relación al RMSE ocurre un patrón similar (mayor precisión al aumentar la muestra y, en ausencia de impacto, para el método O-O).

Tabla 5.5. Media de Sesgo y RMSE en la estimación del tamaño del efecto (ausencia de DIF)

Tamaño de las muestras	Longitud del TAI	*Sin impacto*				*Impacto*			
		O-O	O-M	M-M	IM	O-O	O-M	M-M	IM
		Sesgo							
250-250	10	0.035	0.039	0.039	0.038	0.04	0.041	0.041	0.042
	30	0.037	0.039	0.039	0.038	0.038	0.039	0.039	0.039
500-500	10	0.027	0.03	0.031	0.029	0.031	0.031	0.031	0.033
	30	0.027	0.029	0.029	0.028	0.029	0.03	0.03	0.03
		RMSE							
250-250	10	0.040	0.044	0.044	0.043	0.045	0.046	0.046	0.047
	30	0.042	0.044	0.044	0.043	0.043	0.044	0.044	0.044
500-500	10	0.030	0.034	0.034	0.033	0.035	0.035	0.035	0.036
	30	0.030	0.032	0.032	0.032	0.032	0.033	0.033	0.034

DIF unidireccional

Los resultados del sesgo y RMSE en la estimación de β en presencia de DIF unidireccional se presentan en la tabla 5.6. En este caso se observa que el tamaño del efecto tiende a sobreestimarse cuando el DIF es pequeño [$F(1,69) = 42.86$; $\eta_p^2 = .38$, $\mu_{\beta=0.05} = .010$; $\mu_{\beta=0.10} = .005$]. El aumento del tamaño muestral reduce la sobreestimación [$F(1,69) = 26.07$; $\eta_p^2 = .27$, $\mu_{250-250} = .009$; $\mu_{500-500} = .005$]

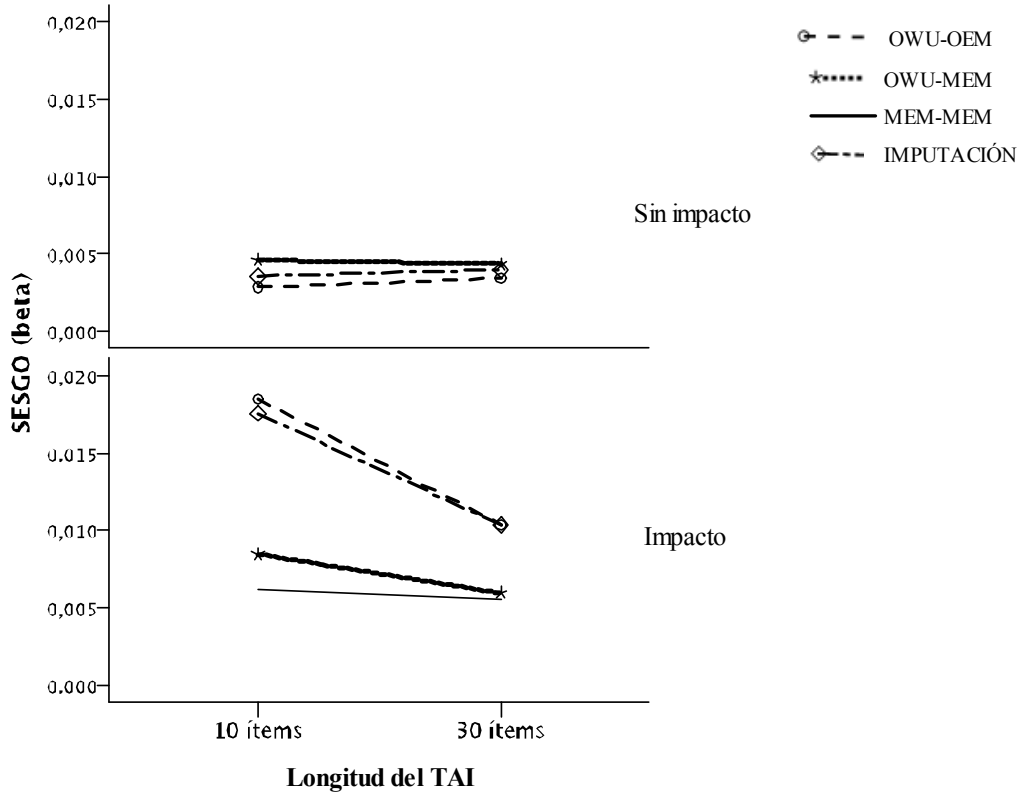
Tabla 5.6. Media de Sesgo y RMSE en la estimación del tamaño del efecto (DIF unidireccional)

β	Tamaño de las muestras	Longitud del TAI	*Sin impacto*				*Impacto*			
			O-O	O-M	M-M	IM	O-O	O-M	M-M	IM
0.05			Sesgo							
	250-250	10	0.007	0.010	0.010	0.009	0.023	0.016	0.014	0.024
		30	0.003	0.004	0.004	0.004	0.018	0.007	0.004	0.017
	500-500	10	0.007	0.008	0.008	0.007	0.014	0.011	0.011	0.015
		30	0.002	0.003	0.003	0.003	0.009	0.004	0.004	0.009
0.10										
	250-250	10	0.002	0.004	0.004	0.003	0.018	0.009	0.007	0.017
		30	-0.001	0.000	0.000	-0.001	0.015	0.003	0.000	0.013
	500-500	10	0.004	0.005	0.005	0.005	0.010	0.006	0.005	0.010
		30	0.001	0.001	0.001	0.001	0.008	0.003	0.002	0.008
0.05			RMSE							
	250-250	10	0.029	0.030	0.030	0.029	0.041	0.034	0.033	0.040
		30	0.028	0.028	0.028	0.028	0.034	0.031	0.031	0.033
	500-500	10	0.022	0.022	0.022	0.021	0.031	0.025	0.025	0.030
		30	0.022	0.022	0.022	0.022	0.025	0.023	0.023	0.025
0.10										
	250-250	10	0.033	0.033	0.033	0.033	0.038	0.034	0.034	0.038
		30	0.031	0.031	0.031	0.031	0.037	0.035	0.035	0.036
	500-500	10	0.024	0.024	0.024	0.024	0.029	0.025	0.024	0.028
		30	0.023	0.023	0.023	0.023	0.026	0.024	0.024	0.026

Los efectos más importantes fueron los efectos del método [$F(3,207) = 90.85$; $\eta_p^2 = .57$], de impacto [$F(1,69) = 72.74$; $\eta_p^2 = .51$] y de la interacción de ambos [$F(3,207) = 155.95$; $\eta_p^2 = .69$]. También hubo un efecto significativo de la interacción

entre método y longitud [$F(3,207) = 15.50$; $\eta_p^2 = .18$] y de la interacción triple entre método, impacto y longitud [$F(3,207) = 25.17$; $\eta_p^2 = .27$]. Esta interacción se ilustra en la figura 5.8.

Figura 5.2. Sesgo en la estimación de β cuando el DIF es unidireccional, como función de método, impacto y longitud del TAI



En ausencia de impacto, apenas existen diferencias entre los métodos y todos presentan un sesgo próximo a cero ($\mu_{O-O} = .003$; $\mu_{O-M} = \mu_{M-M} = \mu_{IM} = .004$). Por el contrario, en presencia de impacto los métodos O-O e IM sobrestiman el tamaño del DIF ($\mu_{O-O} = \mu_{IM} = .014$), mientras que para los métodos que actualizan la distribución posterior (M-M y O-M) se obtiene un sesgo promedio de cero ($\mu_{O-M} = .007$; $\mu_{O-M} = .006$). La interacción triple indica que las diferencias entre los métodos en presencia de impacto son mayores cuando el test es corto.

En relación al RMSE ocurre un patrón similar (mayor precisión al aumentar la muestra y mayor precisión de M-M y O-M en presencia de impacto, especialmente cuando el TAI es corto).

DIF no unidireccional

Los resultados del sesgo en la estimación de β en presencia de DIF no unidireccional se presentan en la tabla 5.7. En este caso hay efectos del tamaño del DIF [$F(1,37) = 209.07$; $\eta_p^2 = .85$] y de la interacción de este con el tamaño muestral [$F(1,37) = 7.94$; $\eta_p^2 = .18$]. Cuando el DIF es grande ocurre una subestimación del DIF independientemente del tamaño muestral ($\mu_{250-250} = -.011; \mu_{500-500} = -.010$). Cuando el DIF es pequeño tiende a producirse una pequeña sobreestimación si la muestra es pequeña ($\mu_{250-250} = .008; \mu_{500-500} = .003$).

Tabla 5.7. Media de Sesgo y RMSE en la estimación del tamaño del efecto (DIF no unidireccional)

β	Tamaño de las muestras	Longitud del TAI	*Sin impacto*				*Impacto*			
			O-O	O-M	M-M	IM	O-O	O-M	M-M	IM
			Sesgo							
0.05										
	250-250	10	0.001	0.007	0.007	0.004	0.009	0.012	0.012	0.012
		30	-0.022	-0.010	-0.010	-0.016	-0.020	-0.010	-0.009	-0.015
	500-500	10	0.005	0.008	0.008	0.007	0.007	0.010	0.010	0.009
		30	-0.016	-0.009	-0.009	-0.012	-0.009	-0.003	-0.003	-0.006
0.10										
	250-250	10	-0.003	0.004	0.004	0.000	0.002	0.007	0.007	0.004
		30	-0.022	-0.008	-0.008	-0.016	-0.019	-0.006	-0.006	-0.014
	500-500	10	0.001	0.004	0.004	0.003	0.001	0.003	0.004	0.002
		30	-0.012	-0.004	-0.004	-0.008	-0.014	-0.005	-0.005	-0.009
			RMSE							
0.05										
	250-250	10	0.024	0.026	0.026	0.024	0.027	0.028	0.028	0.027
		30	0.026	0.027	0.027	0.027	0.026	0.027	0.027	0.026
	500-500	10	0.020	0.022	0.022	0.020	0.020	0.023	0.023	0.020
		30	0.018	0.019	0.019	0.019	0.019	0.020	0.020	0.020
0.10										
	250-250	10	0.033	0.029	0.029	0.030	0.033	0.029	0.029	0.030
		30	0.030	0.028	0.028	0.028	0.028	0.027	0.027	0.027
	500-500	10	0.030	0.024	0.024	0.026	0.030	0.023	0.023	0.026
		30	0.023	0.021	0.021	0.022	0.024	0.021	0.021	0.022

También resultaron efectos significativos del método [$F(3,111) = 49.64$; $\eta_p^2 = .57$] y de la interacción de este con tamaño del DIF [$F(3,111) = 7.86$; $\eta_p^2 = .18$]. Cuando el DIF es grande, I-M y sobre todo O-M subestiman del tamaño del efecto ($\mu_{O-O} = -.017$; $\mu_{IM} = -.012$; $\mu_{O-M} = \mu_{M-M} = -.007$). Cuando el DIF es pequeño, no se

produce subestimación y los resultados de los métodos son muy similares ($\mu_{O-O} = .003$; $\mu_{IM} = .005$; $\mu_{O-M} = \mu_{M-M} = .007$).

En relación al RMSE ocurre un patrón similar. Se encontró mayor precisión al aumentar la muestra y, cuando el DIF era grande, peor precisión de O-M.

Figura 5.3. Sesgo y RMSE en la estimación del parámetro ***a*** del grupo de *referencia* como función del método, impacto y longitud del TAI y tamaño de las muestras

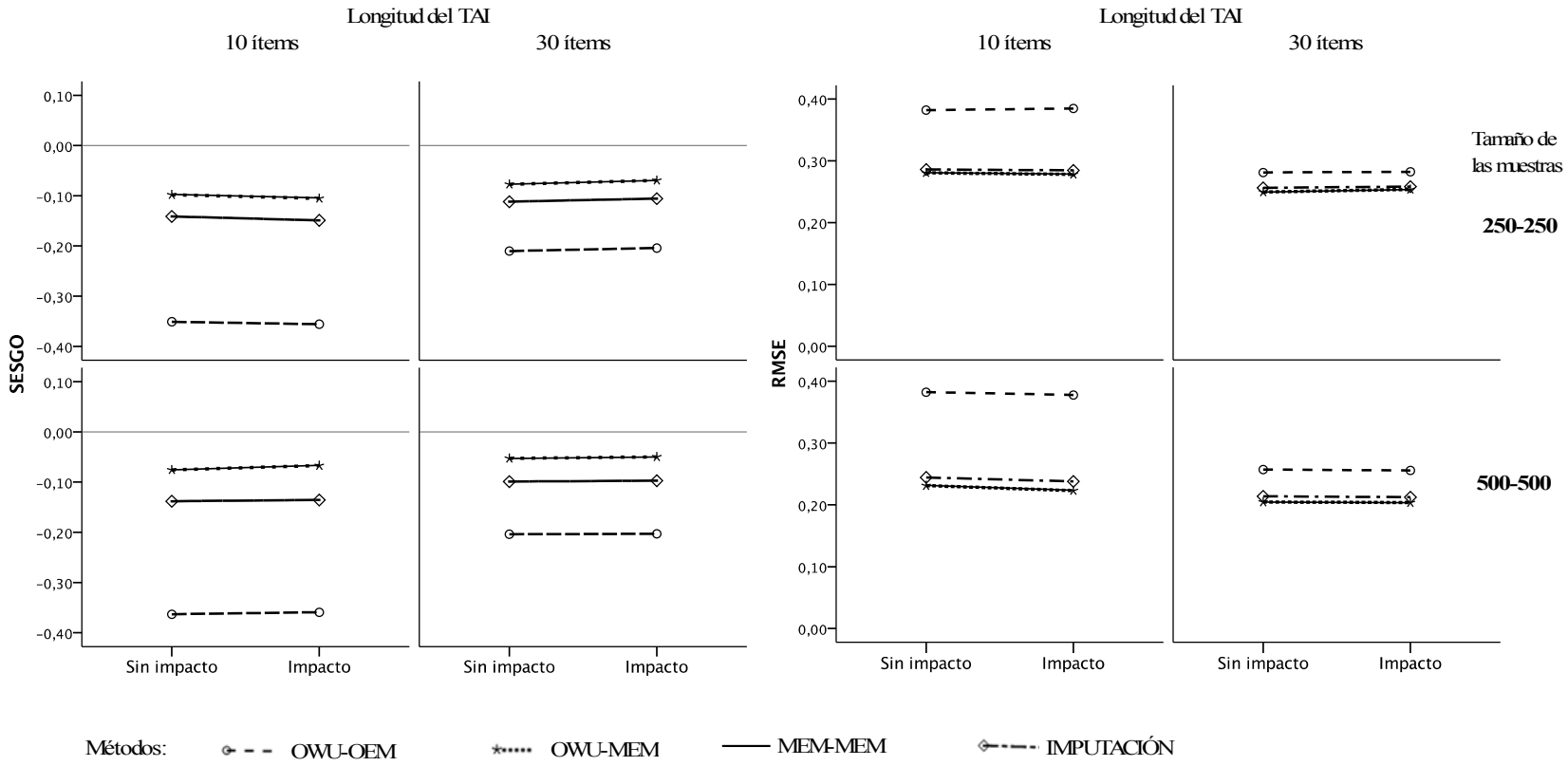


Figura 5.4. Sesgo y RMSE en la estimación del parámetro ***b*** del grupo de *referencia* como función del método, impacto y longitud del TAI y tamaño de las muestras

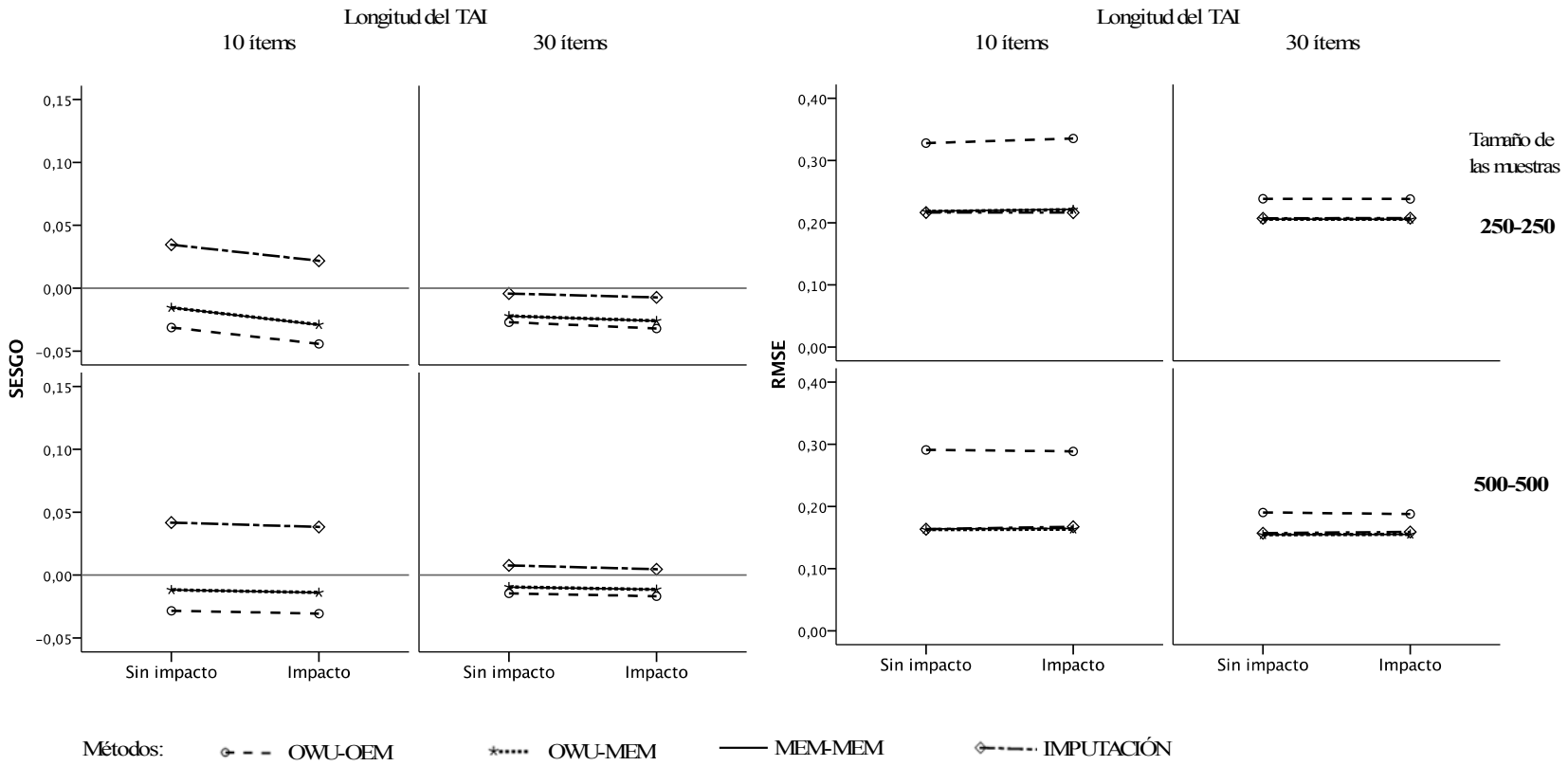


Figura 5.5. Sesgo y RMSE en la estimación del parámetro ***c*** del grupo de *referencia* como función del método, impacto y longitud del TAI y tamaño de las muestras

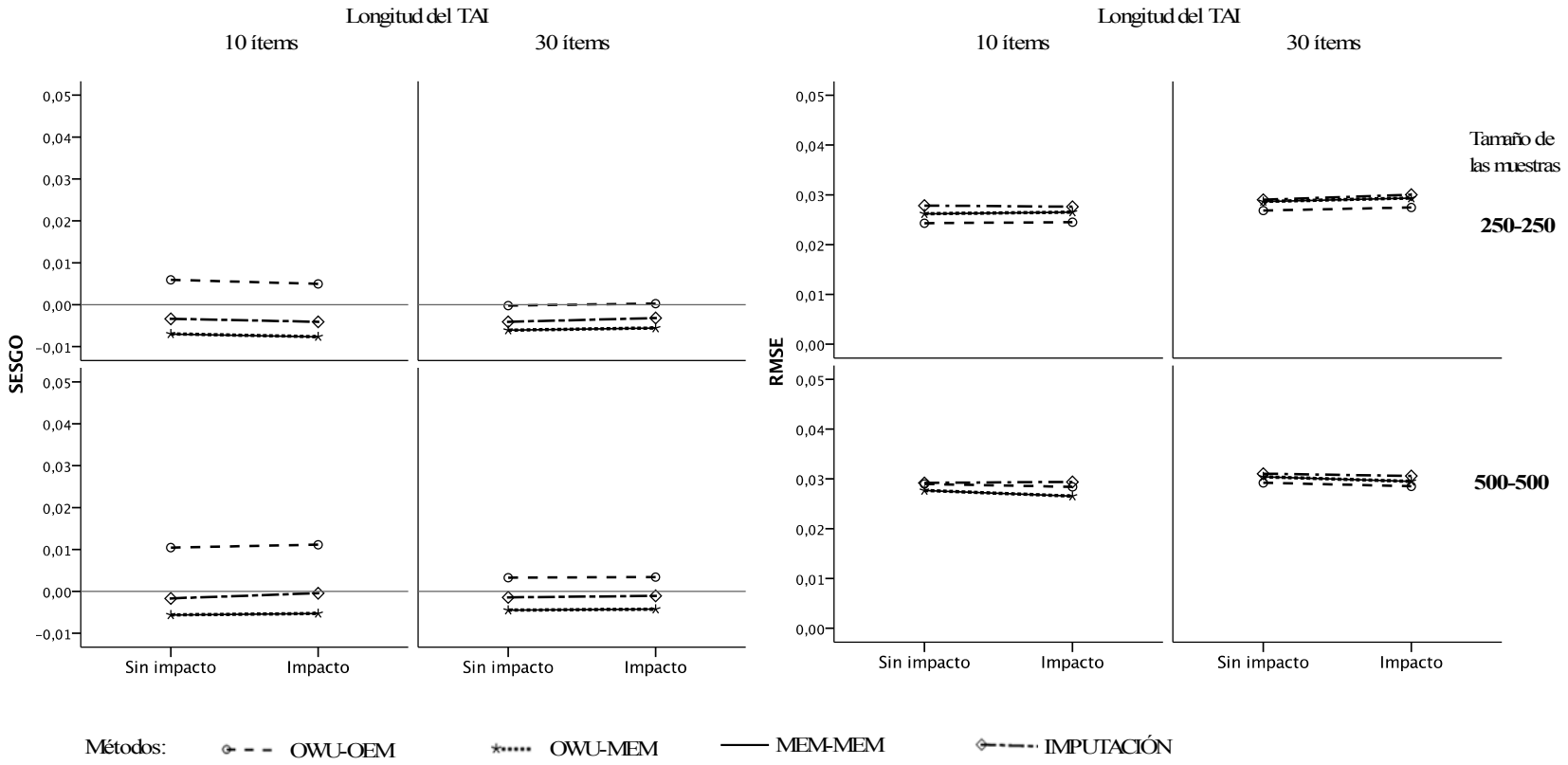


Figura 5.6. Sesgo y RMSE en la estimación del parámetro ***a*** del grupo de *focal* como función del método, impacto y longitud del TAI y tamaño de las muestras

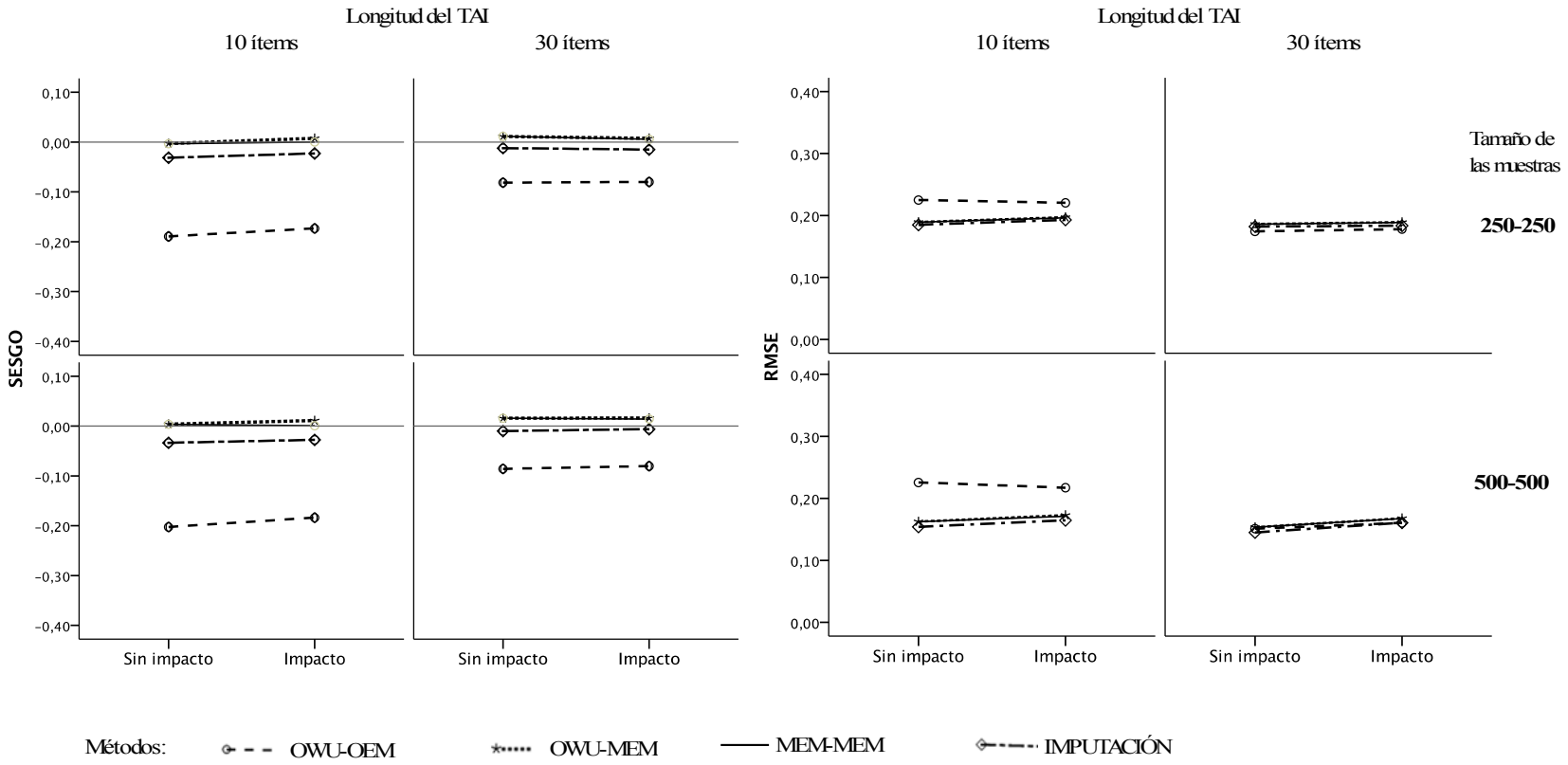


Figura 5.7. Sesgo y RMSE en la estimación del parámetro ***b*** del grupo de *focal* como función del método, impacto y longitud del TAI y tamaño de las muestras

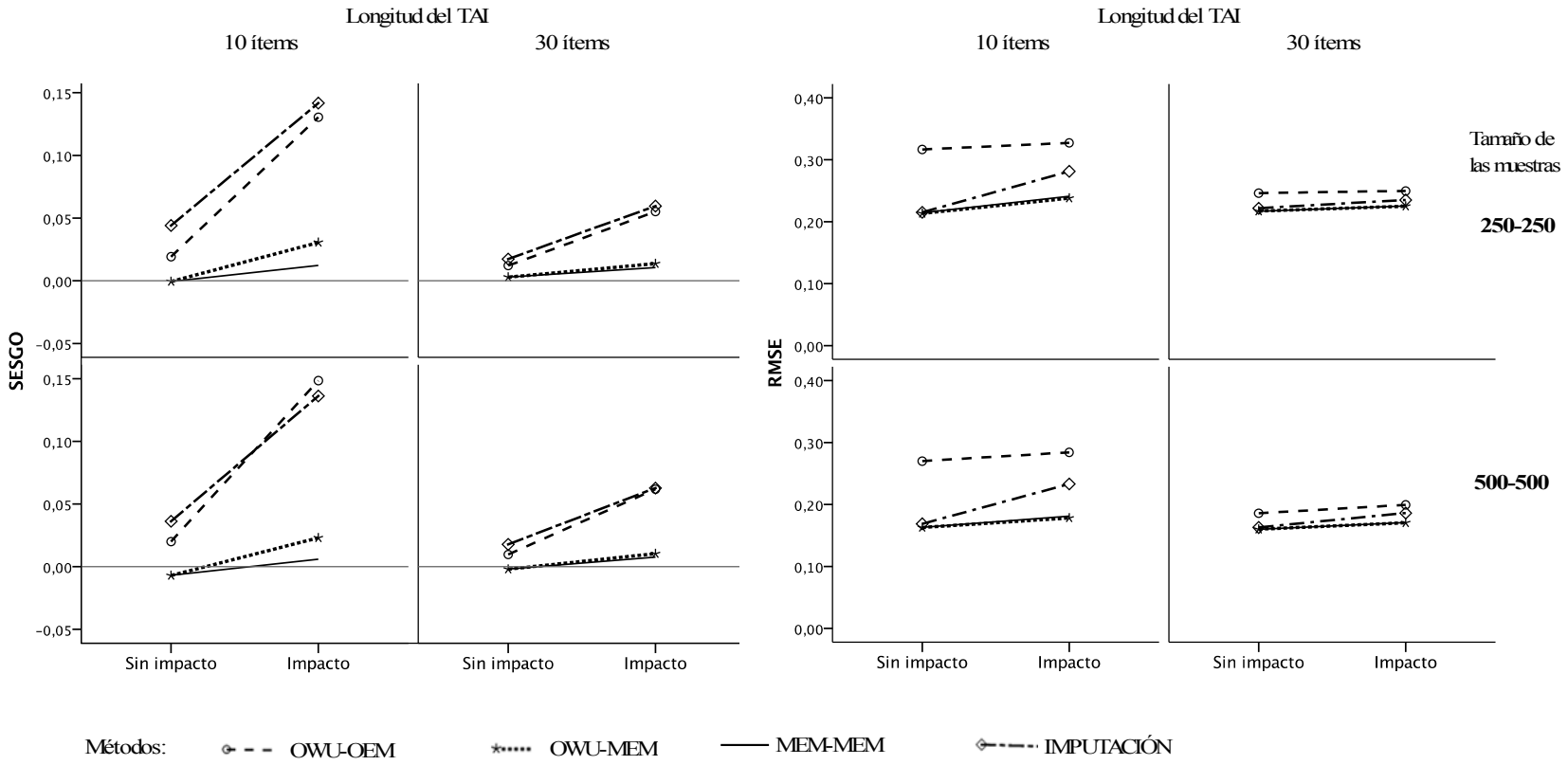
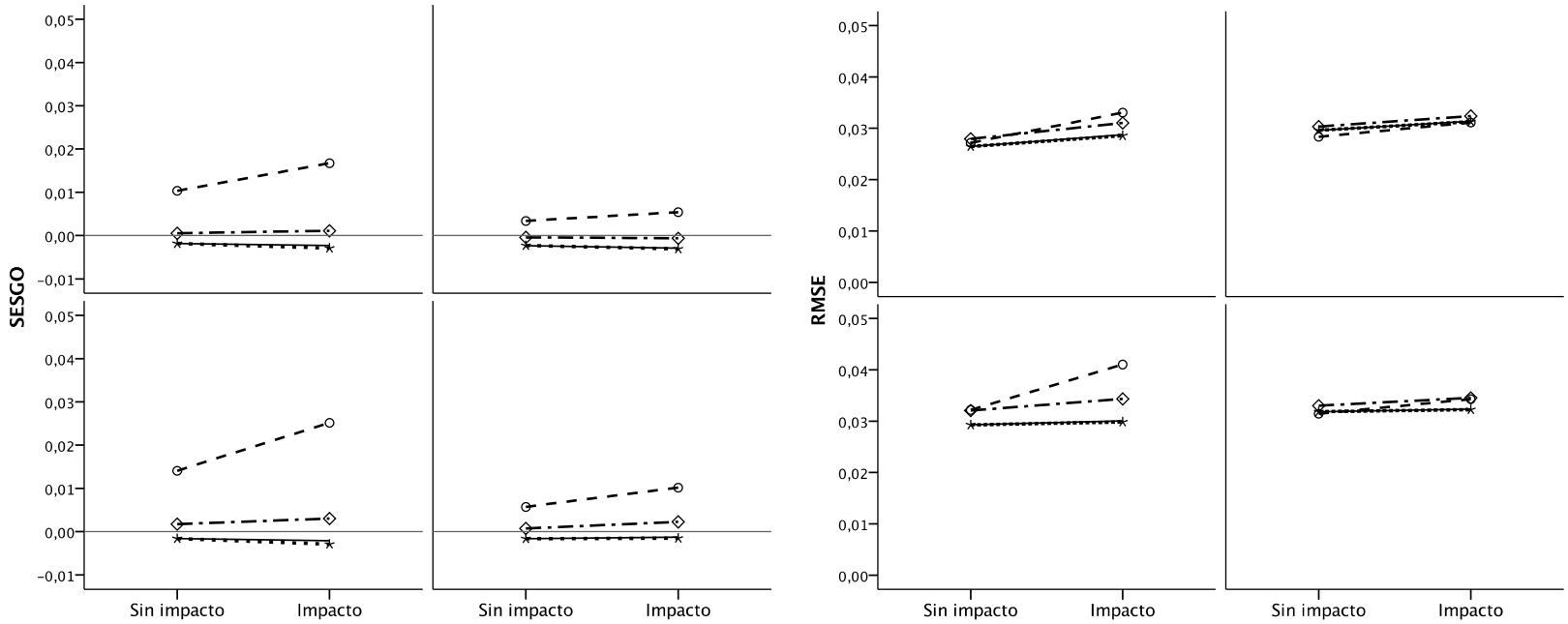


Figura 5.8. Sesgo y RMSE en la estimación del parámetro c del grupo de focal como función del método, impacto y longitud del TAI y tamaño de las muestras



5.7. Discusión y conclusiones

En el presente estudio se evalúo la adaptación de procedimientos de CPF a la prueba de razón de verosimilitudes de la TRI para el análisis del DIF en ítems pretest. Se manipularon diversos factores en la generación de datos y se obtuvieron medidas relacionadas con la identificación del DIF, con la estimación de los parámetros de los ítems y de la distribución de la habilidad, con el fin de valorar la eficiencia relativa de los métodos.

Los resultados sugieren que en todos los casos se produce un adecuado control de la tasa de error tipo I. Las tasas de error tipo I se mantuvieron por debajo del valor nominal independientemente del método utilizado. Estos resultados corroboran diversos estudios en test fijos que señalan que IRT LRT es uno de los procedimientos más eficientes en el control de la tasa de Error tipo I (Cohen, et al., 1996; Finch, 2005; Finch y French, 2007; Kim y Cohen, 1995, 1998; Lopez-Rivas, et al., 2008). De hecho, Lopez-Rivas et al. (2008) informan de tasas de error tipo I no mayores a 0.01 cuando el test de anclaje es óptimo.

Por otro lado, se encontró que, independientemente del método, el tamaño del DIF y el tamaño de las muestras afectan de forma importante a la potencia. La potencia aumenta con el tamaño del DIF y con el tamaño de las muestras. Estos resultados reiteran lo que se ha encontrado en estudios del DIF en test fijos (Finch, 2005; Finch y French, 2007; Stark, et al., 2006; Wang, 2004; Wang y Yeh, 2003), y en tests adaptativos (Lei et al., 2006; Nandakumar y Roussos, 2001, 2004). En el presente estudio se encuentra que se requieren muestras de 1000 examinados para alcanzar valores de potencia adecuados (.8). Además, estos valores sólo se alcanzan en la detección de ítems con DIF grande.

A pesar de que los métodos contrastados muestran un funcionamiento similar se encontraron algunas diferencias en su capacidad para la detección del DIF. La aparición de esas diferencias viene principalmente mediada por dos de los factores manipulados, la longitud del TAI y la presencia de impacto.

En relación a la detección del DIF no unidireccional, el método OWU-OEM tuvo una potencia menor que los otros procedimientos, independientemente del impacto o de la longitud del TAI. Puesto que el DIF no unidireccional se debe a una diferencia en el parámetro *a* entre las CCI de los grupos, una explicación plausible a este resultado

es la poca precisión en la recuperación de dicho parámetro en los procedimientos que utilizan un solo ciclo EM (Ban et al., 2001; Ban et al., 2002; Kim, 2006). Las diferencias en relación a la estimación de β fueron muy pequeñas, pero en general se encontraba una mayor subestimación de β con el método OWU-OEM.

En relación a la detección del DIF unidireccional, el método MWU-MEM con imputación mostró potencias más altas que los otros métodos en presencia de impacto, especialmente cuando el TAI era corto. De forma inesperada la potencia para este método era mayor cuanto menor la longitud del TAI. Estos resultados, contrarios a lo que se esperaba, se deben, en realidad, a una sobreestimación del tamaño del efecto β.

El uso de la imputación hace que se produzca una compresión en la escala (i.e. para ambos grupos se subestima la desviación típica de la distribución del nivel de rasgo) y, en presencia de impacto, se sobreestima la media de la distribución del rasgo en el grupo focal. Estas distorsiones son mayores cuando el TAI es corto y se deben a que, para realizar las imputaciones, se utiliza el nivel de rasgo estimado. Es conocido que el procedimiento MAP produce estimaciones sesgadas hacia la media de la distribución previa y que ese sesgo tiende a corregirse conforme la longitud del test aumenta (Wang y Vispoel, 1998). Este sesgo hace que, en presencia de impacto, las CCI se estimen peor. Se sobreestima la dificultad del ítem en el grupo focal y esto aumenta de manera ficticia la diferencia entre las CCI de los grupos, haciendo que el tamaño del efecto se sobreestime y, por consiguiente, aumente la tasa de detección de DIF.

En relación a la recuperación de los parámetros de los ítems y de la distribución a partir de los otros métodos CPF, los resultados indican que, como era esperado, MWU-MEM presenta estimaciones más precisas independientemente de los factores manipulados. La calidad de la recuperación no depende de la presencia o tamaño del DIF. Así pues, el uso del ítem *pretest* con DIF para la actualización de la distribución previa y la distribución posterior no parece afectar negativamente a los métodos CPF.

En cuanto a OWU-MEM produce una recuperación de los parámetros de los ítems similar a MWU-MEM; las CCI también se recuperan satisfactoriamente; en presencia de impacto, se produce sobreestimación de la media en el grupo focal, pero esto no produce un sesgo importante en la estimación del tamaño del efecto.

En cuanto a OWU-OEM, su funcionamiento es peor que OWU-MEM y MWU-MEM. Lo más notable en relación con la recuperación de los parámetros de los ítems es la dificultad para estimar de manera eficiente el parámetro *a*. Resultados similares han sido señalados en los estudios de CPF en los métodos que sólo emplean un ciclo EM

(Ban, et al., 2001, Ban, et al., 2002; Kim, 2006). Por otro lado, en presencia de impacto, produce distorsiones en la estimación de la distribución del grupo focal, siendo mejor la recuperación de la escala cuando el TAI es más largo (Kim, 2006).

Uno de los criterios fundamentales para la validación de los estudios de DIF es la recuperación del tamaño del DIF. En estas medidas se hace evidente que los métodos más adecuados son OWU-MEM y MWU-MEM, mientras que OWU-OEM y el procedimiento de imputación tienen problemas para recuperar las diferencias entre las CCIs.

Si tomamos en cuenta los resultados sobre los factores y la eficiencia de los métodos podemos concluir que:

1) No es recomendable imputar las respuestas en la matriz de datos generadas por el TAI porque aumenta artificialmente el tamaño del DIF, especialmente en presencia de impacto y si el TAI es corto.
2) Por el contrario, se considera que el resto de los procedimientos CPF, especialmente el método MWU-MEM, son más eficientes. La calibración no se ve afectada por el desajuste que aporta un ítem con DIF.
3) La decisión de descartar un ítem por la presencia de DIF debe estar basada tanto en una medida de significación estadística como en una medida que evalúe la magnitud del DIF.

Así pues, el presente estudio muestra que los procedimientos de CPF son métodos prácticos y fiables que permiten detectar el DIF en ítems pretest que son aplicados en procedimientos de calibración online. Las estimaciones, sobre todo del método MWU-MEM, fueron estables aún con tamaños de muestra pequeños, lo que permite sugerir el uso de esta técnica en las primeras fases de estudio de las propiedades psicométricas de ítem pretest.

También se ha propuesto una modificación del método de imputación de Lei, et al. (2006), reemplazando el procedimiento CPL por un método CPF. Aunque el funcionamiento es adecuado en la mayor parte de las condiciones, nuestros resultados ilustran que el uso de la imputación parece desaconsejable en presencia de impacto cuando el TAI es corto, pues puede dar lugar a distorsiones en la estimación del tamaño del efecto.

5.8. Referencias

Ban, J.C., Hanson, B.A., Wang, T., Yi, Q., y Harris, D.J. (2001). A comparative study of on-line pretest-item calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, *38*, 191-212.

Ban, J.C., Hanson, B.A., Yi, Q., Harris, D.J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement*, *39*, 207-218.

Bock, R.D., y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: *Application of an EM algorithm. Psychometrika*, *46*, 443-459.

Clauser, B., y Mazor, K., (1998) Using statistical procedures to identify differentially funcioning test items. *Instructional topics in educational measurement*. National Council on Measurement in Education, 31-44.

Cohen, A., Kim. S. y Wollack, J. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15-26.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2ª ed. Hillsdale, NJ. Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.

Cuevas, L., Abad, F., Olea, J., Barrada, J.R. y Garrido L. (2010). *CATSIM 1.0*. software presented at the IV European Congress of Methodology. Potsdam, Germany.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295.

Finch, W. H., y French, B. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*, 565-582.

Folk, V. y Golub-Smith, M. (1996). *Calibration of on-line pretest data using BILOG*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

French, B., y Finch, H. (2008). Multigroup confirmatory factor analysis: locating the invariant referent sets. *Structural Equation Modeling, 15*, 96-113.

Hanson, B. A. (2002). *IRT Command Language (Version 0.020301)*. Monterey, CA: Author. (Available at http://sourceforge.net/projects/ssm)

Harmes, J., Kromrey, J. y Parshall, C.(2001). *Online item parameter recalibration: application of missing data treatments to overcome the effects of sparse data conditions in a computerized adaptive version of the MCAT.* Report submitted to the Association of American Medical Colleges. Section for the MCAT. 1-27.

Haynie, K.A. y Way, W.D. (1995). *An investigation of item calibration procedure for a computerized licensure examination.* Paper presented at a symposium entitled

Computer Adaptive Testing, at the annual meeting of the National Council on Measurement in Education, San Francisco.

Hsu, Y., Thompson, T.D., y Chen, W.H. (1998). *CAT item calibration*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*, 355-381.

Kim, S., y Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8,* 291-312.

Kim, S., y Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22,* 345-355.

Lei, P.-W., Chen, S.-Y. y Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*, 245–264.

Lopez-Rivas, G., Stark, S., y Chernyshenko, O. (2008). The effects of referent item parameters on Differential Item Functioning detection using the Free Baseline Likelihood Ratio Test. *Applied Psychological Measurement, 33*, 251-265.

Mills, C. N., y Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing, *Applied Measurement in Education, 9*, 287-304.

Nandakumar, R. y Roussos, L. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests* (Research report). Newtown, PA: Law School Admission Council.

Nandakumar, R. y Roussos, L. A. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29*, 177–200.

Parshall, C.G., Kromrey, J.D., Harmes, J.C., y Sentovich, C. (2001). *Nearest neighbors, simple strata, and probabilistic parameters: An empirical comparison of methods for item exposure control in CATs.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle.

Revuelta, J., y Ponsoda, V. (1998). A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement, 35,* 311-27.

Segall, D. (2003*). Calibrating pools and online pretest items using MCMC methods*. Paper presented at the annual meeting of the NCME. Chicago. 1-9.

Stark, S., Chernyshenko, O. S., y Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1291-1306.

Stocking, M. (1988). *Scale Drift in online calibration*. (Research Report ETS-RR-88-28-ONR). Princeton, NJ:ETS.

Thissen, D. (2001). *IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software and manual*]. Chapel Hill: L.L. Thurstone Psychometric Laboratory, University of North Carolina.

Wainer, H. (1990). Model-based standardized measurement of an item´s differential impact. En P. W. Holland y H. Wainer (Eds.). *Differential item functioning (pp. 123-153)*. Mahwah, NJ: Erlbaum.

Wainer, H., Bradlow, E., y Wang, X. (2010). Detecting DIF: Many paths to salvation. *Journal of educational and behavioral statistic, 35*, 489-493.

Wang, T. y Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.

Wang, W. (2004). Effects of anchor item methods on detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261.

Wang, W., y Yeh, Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Woods, C. (2008a). Empirical selection of anchors items. *Applied Psychological Measurement, 33*, 42-57.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. En W. J. van der Linden y C. A. W. Glas (Eds.), Computerized Adaptive Testing. *Theory and Practice (págs. 221-243)*. Boston, MA: Kluwer Academic Publishers.

Zwick, R. (2007). The investigation of Differential Item Functioning in adaptive test. En W. J. van der Linden y C. A. W. Glas (Eds.), *Elements of adaptive testing. (págs. 331-352)*. New York: Springer.

Capítulo 6

Estudio 2. Evaluación del funcionamiento diferencial en ítems aplicados mediante un test adaptativo informatizado: IRT LRT y CATSIB

6.1. Abstract

El uso cada vez más extendido de los test adaptativos informatizados (TAIs) ha mostrado la necesidad de adecuar las técnicas que evalúan el funcionamiento diferencial del ítem (DIF) a este tipo de tests. En el presente trabajo se comparó el test de razón de verosimilitudes de la TRI (IRT LRT) y el CATSIB (con corrección en la puntuación y sin corrección) para la detección del DIF en ítems aplicados adaptativamente. De un banco de 300 ítems, se eligieron los 30 con mayor tasa de exposición para el análisis del DIF. Para simular DIF se disminuyó en 0.4 el valor del parámetro de dificultad en el grupo focal en 15 de los 30 ítems que se analizaban. La efectividad de los métodos se evaluó en términos de la tasa de error Tipo I en ausencia de DIF (banco no contaminado) y en presencia de otros ítems con DIF (banco contaminado), de potencia y de recuperación del tamaño del DIF. Las variables manipuladas fueron presencia de contaminación en el banco, impacto y tamaño de las muestras. IRT LRT controló el error tipo I satisfactoriamente en todas las condiciones, mientras que con CATSIB las tasas de error tipo I en el banco contaminado se situaron por encima del valor nominal, lo que fue más evidente para el método sin corrección y con presencia de impacto. El RMSE para la recuperación del tamaño del DIF fue menor para IRT LRT. Se recomienda el uso del IRT LRT para el análisis de DIF en TAIs.

Palabras clave: funcionamiento diferencial del ítem, test adaptativo informatizado, prueba de razón de verosimilitudes de la teoría de respuesta al ítem, CATSIB

En la última década, muchos de los programas de evaluación a gran escala que aplican tests adaptativos informatizados (TAIs) se han consolidado (Cheng, 2009; Kingsbury, 2009; Walter, Postma, McHugh, Rush, Coyle, Strong, Sharpe, 2007) y uno de los retos para mantener su viabilidad es desarrollar procedimientos eficientes relacionados con el mantenimiento y renovación del banco de ítems (Abad, Olea, Aguado, Ponsoda y Barrada, 2010; Pommerich, Segall y Moreno, 2009). Por ello, se espera que el desarrollo de procedimientos para el estudio de la validez de los ítems operativos de un modo automatizado sea uno de los desafios importantes en los próximos años (Embretson, 2004).

La recalibración se recomienda periódicamente para estudiar posibles cambios en los parámetros debidos a: (a) una posible filtración al público del contenido del ítem; (b) cambios en los contenidos del constructo (DeMars, 2004; Glass, 2007; Wells, Subkoviak y Serlin, 2002; Mills y Stocking, 1996); (c) cambios en el modo de administración del test, cuando el programa de TAI se aplica usando los parámetros estimados de una aplicación previa en lápiz y papel (Harmes, Kromrey y Parshall, 2001); o (d) la aplicación en diferentes poblaciones (Hart, Deutsher, Crane y Wang, 2009; Walter, et al., 2007; Zwick, 2007). A pesar de su importancia, pocos estudios se han dirigido a la recalibración y el análisis del DIF en ítems operativos (p.ej. Kingsbury y Houser, 1993; Haynie y Way, 1995; Zwick, Thayer y Wingersky, 1993, 1994a).

La recalibración de los parámetros de los ítems operativos ha de afrontar dos problemas. Primero, puesto que el tamaño del test es menor que el tamaño del banco (no menos de 6 –Way, 1998– u 12 veces –Stocking, 1994– menor), la mayor parte de la matriz de datos contiene valores faltantes. Segundo, dado que el TAI está diseñado para maximizar la eficiencia, los examinados responden a aquellos ítems que son más informativos para su nivel de habilidad estimado. Esto produce que: (a) la distribución del nivel de habilidad de quienes responden a los ítems cambie de ítem a ítem; y (b) la variabilidad del nivel de rasgo de quienes responden a los ítems sea marcadamente menor que la variabilidad muestral (Harmes et al., 2001; Haynie y Way, 1995; Ito y Sykes, 1994).

Estos problemas quedan satisfactoriamente resueltos con los métodos de calibración con parámetros fijos (CPF; Ban et al., 2002; Ban et al., 2001; Kim, 2006). En la CPF se fijan los parámetros de ciertos ítems a sus valores previamente calculados

mientras que se estiman los parámetros de otros ítems. Así, por ejemplo, en los diseños de calibración online, se fijan los parámetros de los ítems operativos y se estiman los de los ítems pretest.

En el caso de recalibración de ítems operativos, nuestra propuesta es que los ítems con parámetros fijos sean todos los ítems operativos menos uno, aquel que está siendo recalibrado. Para ello, los responsables del programa de evaluación fijan de antemano un cierto número de presentaciones del ítem que, al ser alcanzado, supone la recalibración del ítem. El análisis del ítem se realiza con la submuestra de examinados a la que se ha administrado el ítem a analizar. De este modo: (a) la frecuencia de análisis viene condicionada por el uso del ítem; (b) la calidad de las estimaciones de los ítems es similar, al estar basada en idénticos tamaños muestrales; y (c) se usa toda la información disponible sobre los parámetros de los ítems del banco. Un posible inconveniente es que el análisis puede verse afectado por la presencia de desajustes en ítems distintos al que está siendo analizado.

Las diferentes situaciones que hacen recomendable la recalibración de ítems han sido tratadas como campos distintos en el estudio de la validez. Sin embargo, con algunos cambios se pueden utilizar los procedimientos de DIF para estudiar los diferentes escenarios (p.ej., Abad et al., 2010; DeMars, 2004; Veldkamp y van der Linden, 2007). Las técnicas de DIF pueden servir para comprobar la ausencia de DIF entre dos grupos en el modo de aplicación adaptativo, pero también para comprobar si los parámetros de los ítems operativos se han modificado a través del tiempo o a través de las condiciones de aplicación.

En los últimos años se han adaptado diferentes métodos para el análisis de DIF en TAIs, como el test de razón de verosimilitudes de la TRI (IRT LRT; Lei, Chen y Yu, 2006), el CATSIB (Nandakumar y Roussos, 2001, 2004) y la regresión logística (Lei et al., 2006). El uso de un procedimiento basado en la TRI resulta la opción más natural en este contexto, pues ésta se aplica a lo largo de todo el proceso de elaboración del TAI (p.ej., la calibración del banco de ítems, la selección de los ítems mediante el algoritmo adaptativo o la puntuación de las respuestas) y porque permite conocer directamente como varían los parámetros de los ítems de una condición a otra, de un momento a otro o de un grupo a otro. En el presente estudio se compara la técnica IRT LRT con el CATSIB, tomando este último como línea base, ya que, entre los procedimientos no basados en la TRI, CATSIB es el más prometedor en la detección del DIF unidireccional (ver capítulo 3).

6.2. IRT LRT

Asumamos dos grupos de evaluados, referencia y focal (g:{R, F}). IRT LRT es un método paramétrico en el que se evalúa la presencia de DIF a través de la comparación del ajuste en modelos de TRI anidados (Thissen, Steinberg y Gerrard, 1986; Thissen, Steinberg y Wainer, 1988): un modelo en el que todos los parámetros del ítem estudiado se restringen a ser iguales entre los grupos (modelo compacto) y un modelo en el que se permite que dichos parámetros varíen (modelo aumentado). El estadístico de prueba es:

$$G^2 = -2\ln L_C - (-2\ln L_A) \tag{6.1}$$

donde L_C y L_A son las verosimilitudes del modelo compacto y aumentado, respectivamente. El estadístico resultante se distribuye aproximadamente como χ^2 con un número de grados de libertad igual a la diferencia de parámetros entre los modelos. Una diferencia estadística indica la presencia de DIF. Dos serían las ventajas principales del método IRT LRT. Por un lado, el uso del modelo psicométrico en el que están fundamentados los TAIs, la TRI. Por otro, la integración, en el mismo procedimiento, de la recalibración y análisis del DIF.

En el estudio de Lei et al. (2006) el análisis de DIF se centra en ítems pretest. La técnica de calibración empleada es la de calibración con parámetros libres, por la que los parámetros de los ítems operativos pueden cambiar entre grupos en la nueva calibración. González-Betanzos, Abad y Barrada (este volumen) han señalado las ventajas, tanto teóricas como aplicadas, de emplear la CPF en el análisis de DIF en ítems pretest. Idénticas razones sirven para justificar la CPF para el caso de ítems operativos. En ese caso, los parámetros de los ítems operativos se fijan a sus valores originales tanto en el modelo compacto como en el modelo aumentado. Los parámetros de distribución del nivel de rasgo (media y desviación típica del rasgo en ambos grupos) se estiman para el modelo aumentado; esos parámetros estimados se asumen en el modelo compacto. En relación a qué método CPF aplicar, González-Betanzos et al. recomiendan el uso del método MWU-MEM.

6.3. CATSIB

CATSIB es un método no paramétrico basado en la estimación del tamaño del DIF, definido como:

$$\beta_{j(CATSIB)} = \int \left(P_{jR}(u_j = 1 \mid \theta, g = R) - P_{jF}(u_j = 1 \mid \theta, g = F) \right) f_j(\theta) d\theta \qquad (6.2)$$

donde $P_j(u_j = 1 \mid \theta, g = R)$ y $P_j(u_j = 1 \mid \theta, g = F)$ marcan la probabilidad de acierto al ítem *j* condicionada al nivel de habilidad (que actúa como variable de igualación) y al grupo de pertenencia del examinado y $f_j(\theta)$ es la función de densidad del grupo total que responde al ítem *j* (Nandakumar y Roussos; 2001). La presencia de DIF supone que la diferencia promedio entre las proporciones condicionadas de acierto de ambos grupos es distinta de 0.

Para la estimación muestral de $\beta_{j(CATSIB)}$ es necesario contar con estimaciones del nivel de habilidad y de las probabilidades de acierto:

- Estimaciones del nivel de habilidad: Una idea obvia es emplear como variable de igualación el nivel de rasgo estimado con el TAI, $\hat{\theta}_{TAI}$. Sin embargo, cuando la distribución del nivel de habilidad difiere entre el grupo de referencia y el focal (p.ej., hay impacto), el valor esperado de habilidad real dado el nivel estimado difiere a través de los grupos. Por ello, se emplea una nueva variable de igualación, $\hat{\theta}^{*}_{TAI}$, calculada mediante la corrección por regresión (Nandakumar y Roussos, 2004).
- Probabilidades de acierto: Estas probabilidades de acierto condicionadas a nivel de habilidad se calculan mediante la discretización del continuo del nivel de habilidad, $\hat{\theta}_{TAI}$, en *K* niveles. La probabilidad ya no se calcula condicionada al nivel de habilidad, sino al intervalo de habilidad. Así, para el grupo de referencia, la probabilidad de acierto en el nivel *k* se obtiene como

$$\hat{P}_{jkR} = \frac{u_{jkR}}{n_{jkR}} \qquad (6.3)$$

donde u_{jkR} es el número de evaluados que aciertan el ítem *j* y están en el nivel *k*; n_{jkR} es el número de examinados que responden al ítem *j* y están en el nivel *k*, lo mismo se aplica para el grupo focal.

El estimador del tamaño del efecto es:

$$\hat{\beta}_{j(CATSIB-SC)} = \sum_{k=1}^{K} (\hat{P}_{jkR} - \hat{P}_{jkF}) f_k \tag{6.4}$$

Donde f_k es la proporción de gente en el nivel *k*:

$$f_k = \left(\frac{n_{jkR} + n_{jkF}}{\sum_{k=1}^{K} (n_{jkR} + n_{jkF})} \right) \tag{6.5}$$

Análogamente, si se utiliza la puntuación corregida, $\hat{\theta}^*_{TAI}$, para agrupar a los sujetos, se calcularía como:

$$\hat{\beta}_{j(CATSIB-C)} = \sum_{k=1}^{K} (\hat{P}^*_{jkR} - \hat{P}^*_{jkF}) f^*_k \tag{6.6}$$

Nandakumar y Roussos (2001, 2004) detallan cómo calcular $\hat{\theta}^*_{TAI}$, y el error típico de $\hat{\beta}_{CATSIB}$ necesario para el contraste de hipótesis. Por tanto, CATSIB proporciona tanto una prueba estadística de la presencia de DIF como un estimador del tamaño del efecto.

En principio no es claro qué variable de igualación ($\hat{\theta}^*_{TAI}$ o $\hat{\theta}_{TAI}$) puede ser mejor en la detección del DIF en ítems operativos. En el análisis de ítems pretest, se ha encontrado que $\hat{\theta}^*_{TAI}$ es preferible (Nandakumar y Roussos, 2001, 2004) pero no está claro que este resultado sea generalizable al análisis de ítems operativos. En este caso, $\hat{\theta}^*_{TAI}$ incluye al propio ítem analizado y existen dudas en relación a aplicar la corrección en este caso. Por ejemplo, en el análisis de ítems que siguen el modelo de Rasch, Holland y Thayer (1988) muestran que el sesgo del estadístico Δ_{MH} en presencia de impacto, desaparece cuando se utiliza la puntuación total, incluyendo el ítem analizado, por lo que sería innecesario establecer una corrección por regresión adicional. También es conocido que incluir el ítem no es suficiente para corregir el sesgo cuando los ítems no siguen el modelo de Rasch pero se desconoce en qué medida es apropiado aplicar ambas correcciones simultáneamente. Al respecto, Walker, Beretvas y Ackerman (2001) realizaron un estudio que comparaba SIBTEST y CATSIB, en el contexto de los tests fijos cuando se incluye el ítem analizado en la variable de igualación. Los

evaluados simulados respondían a un test de L&P de 30 ítems, tres de los cuales tenían DIF. En su estudio, los autores compararon los resultados obtenidos utilizando como variables de igualación la puntuación total *X* (SIBTEST), la estimación en el test fijo $\hat{\theta}_{fijo}$ (CATSIB-C) y la estimación corregida $\hat{\theta}^{*}_{fijo}$ (CATSIB-SC), que se calculaban a partir de todos los ítems. Los resultados mostraron que las tasas de error tipo I eran mayores cuando se utilizó la versión corregida. Los autores sugieren no utilizar la corrección en estos casos. Sin embargo, no es claro si estos resultados sean generalizables al contexto en el que la adaptación del ítem analizado es adaptativa.

6.4. Tamaño del DIF en TAIs

Las pruebas de detección de DIF, como cualquier otra prueba estadística, son sensibles al tamaño muestral. Por ello, es común que las decisiones sobre los ítems analizados no se basen únicamente en el resultado del contraste de hipótesis, sino también en el tamaño del DIF estimado (Camilli y Sherapard, 1994; Jodoin y Gierl, 2001; Roussos y Stout, 1996; Steinberg y Thissen, 2006). Como hemos visto, CATSIB ofrece una medida del tamaño del efecto del DIF.

Una medida análoga a la anterior en el contexto de la TRI puede obtenerse como:

$$\hat{\beta}_{j(IRTLR)|TAI} = \sum_{k=1}^{K} (P_{jR}(u_j = 1 \mid \theta_k, \hat{\boldsymbol{\Delta}}_{jR}, g = R) - P_{jF}(u_j = 1 \mid \theta_k, \hat{\boldsymbol{\Delta}}_{jF}, g = F)) f_j(\theta_k \mid \hat{\boldsymbol{\pi}}_{(j)}) \quad (6.7)$$

donde $f_j(\theta_k \mid \hat{\boldsymbol{\pi}}_{(j)})$ indica la probabilidad de θ_k según una función de distribución normal discretizada. $\hat{\boldsymbol{\pi}}_{(j)}$ es el vector con la media y la desviación típica del grupo total que ha respondido al ítem *j* ($\hat{\boldsymbol{\mu}}_{(j)}, \hat{\boldsymbol{\sigma}}_{(j)}$). Es importante señalar que, en el contexto de los TAIs, $f_j(\theta_k \mid \hat{\boldsymbol{\pi}}_{(j)})$ varía de ítem a ítem, ya que cada ítem es respondido por distintas personas. Por tanto, las medidas de tamaño del efecto pueden diferir de las obtenidas en un contexto no adaptativo. En el contexto no adaptativo, $\hat{\beta}_{j(IRTLR)}$ se obtendría como:

$$\hat{\beta}_{j(IRTLR)|fijo} = \sum_{k=1}^{K} (P_j(u_j = 1 \mid \theta_k, \hat{\boldsymbol{\Delta}}_{jR}, g = R) - P_j(u_j = 1 \mid \theta_k, \hat{\boldsymbol{\Delta}}_{jF}, g = F)) f(\theta_k \mid \hat{\boldsymbol{\pi}}) \quad (6.8)$$

donde $f(\theta_k \mid \hat{\boldsymbol{\pi}})$ es la distribución del grupo al que se le aplica el test fijo, que es igual para todos los ítems. Esto indica la relevancia de reconsiderar la medida de tamaño del efecto del DIF en el contexto adaptativo. Por ejemplo, un ítem difícil con DIF

unidireccional puede tener una valor de $\hat{\beta}_{j(IRTLR)|fijo}$ pequeño, al aplicarse en una muestra de calibración en la que el rasgo siga una distribución normal estándar. Sin embargo, al convertirse en ítem operativo, será aplicado a examinados con mayor nivel de habilidad y el valor de $\hat{\beta}_{j(IRTLR)TAI}$ será mayor.

El propósito de este estudio es comparar la ejecución de IRT LRT, CATSIB, con corrección (CATSIB-C) y sin corrección (CATSIB-SC), para el análisis del DIF en ítems operativos. Se va a comparar la tasa de error Tipo I cuando ninguno de los ítems del banco tiene DIF (banco no contaminado) y cuando un porcentaje de ellos presenta DIF (banco contaminado). En éste último se calculará la potencia para los ítems con DIF. Finalmente, se estudia la recuperación del tamaño del DIF según los distintos procedimientos. Se espera que la TRI proporcione mejores resultados que CATSIB-SC y CATSIB-C, ya que estos dos últimos procedimientos tienen limitaciones en la corrección del sesgo producido en presencia de impacto. Incluir el ítem en el test (CATSIB-SC) supone una corrección insuficiente del sesgo cuando las respuestas a los ítems no siguen el modelo de Rasch. Incluir el ítem y establecer adicionalmente la corrección (CATSIB-C) puede suponer una sobre-corrección del sesgo tal y como encuentran Walker et al. (2001) en el contexto de los tests fijos. Por otro lado, se espera que los procedimientos de CPF, que eliminan la necesidad de recalibrar los parámetros de los ítems operativos en presencia de DIF, haga más robusto el procedimiento IRT LRT a la presencia de contaminación.

6.5. Método

6.5.1. Condiciones de aplicación del TAI

Para la simulación de aplicaciones adaptativas de una prueba se empleó el programa CATSIM 1.0 (Cuevas, Abad, Olea, Barrada y Garrido, 2010). Tanto para el grupo de referencia como el grupo focal se simularon 10,000 examinados por grupo. A cada examinado se le aplicaba un TAI de longitud fija. El nivel de rasgo inicial era asignado al azar dentro del intervalo (–0.4, 0.4). Se empleó estimación MAP con una distribución normal estándar como distribución a priori. Con el fin de reducir el problema de la infraexposición de parte del banco, se empleó como regla de selección

de ítems el método progresivo (Revuelta y Ponsoda, 1998) según el cual la selección de ítems tiene un elevado componente aleatorio al comienzo del test y la importancia de la información de Fisher en la selección se va incrementando según avanza el test. Se utilizó el método Sympson-Hetter (Sympson y Hetter, 1985) para el control de la exposición de la tasa máxima de exposición con una tasa límite de .25.

6.5.2. Banco de ítems

Se generó un banco de 300 ítems. Inicialmente, los parámetros de los ítems operativos eran iguales para el grupo de referencia y para el grupo focal. Las distribuciones de los parámetros se aproximan a las de los ítems del Law School Admission Test y se generaron siguiendo la información proporcionada por Nandakumar y Roussos (2004). Para ítems con parámetro *b* menor o igual a –1, el parámetro *a* seguía una distribución lognormal(–0.357, 0.25) dentro del rango [0.4, 1.1]. Para el resto de ítems, la distribución del parámetro *a* es lognormal(–0.223, 0.34) dentro del rango [0.4, 1.7]. El parámetro *a* promedio dentro del banco era de 0.81. El parámetro *b* seguía una distribución *N*(0, 1) dentro del rango [–3, 3]. El parámetro *c* seguía una distribución uniforme entre [0.12, 0.22].

6.5.3. Ítems analizados

Se realizó una primera simulación de respuestas para un grupo de evaluados con distribución del nivel de rasgo normal estándar, *N*(0, 1). Los ítems del banco fueron ordenados según su tasa de exposición de menor a mayor. Los ítems en el rango 271-300, un 10% del banco, fueron los ítems en los que centramos nuestros análisis. Esta opción parece conveniente, puesto que, a mayor tasa de exposición, más riesgo de desajuste de un ítem y más efectos negativos de la presencia no detectada de un ítem con DIF en la calidad de un programa de evaluación. Los parámetros de estos ítems pueden verse en la Tabla 6.1. Como cabía esperar, el parámetro *a* de los ítems con mayor exposición es mayor que el promedio del parámetro *a* en el banco. En las condiciones con presencia de DIF éste se simuló en los ítems del 286 al 300. Para ello, el parámetro *b* del grupo focal fue el resultado de restar 0.40 al parámetro *b* del grupo de referencia. La mayor facilidad para el grupo focal quería representar una posible

filtración al público del contenido de estos ítems. En la Tabla 6.1 se informa también del tamaño del DIF promedio en las condiciones en que éste estuvo presente, $\overline{\beta}_{j(IRTLR)|TAI}$. El promedio se obtenía a través de las réplicas, utilizando los parámetros reales de los ítems y asumiendo una distribución normal de la habilidad, con la media y la desviación típica reales para el grupo de evaluados que respondieron a dicho ítem.

Tabla 6.1. Parámetros de los ítems estudiados en el Banco contaminado

	SIN DIF				**DIF UNIDIRECCIONAL**					
Item	$a_R=a_F$	$b_R=b_F$	$c_R=c_F$	**Item**	$a_R=a_F$	b_R	b_F	$c_R=c_F$	$\overline{\beta}$	Sβ
271	1.64	0.27	0.20	**286**	1.59	0.07	-0.33	0.15	-0.14	0.0021
272	1.41	1.47	0.14	**287**	1.46	-0.25	0.19	0.14	-0.13	0.0017
273	1.37	-0.08	0.19	**288**	1.34	0.86	-0.65	0.13	-0.13	0.0017
274	1.33	1.44	0.14	**289**	1.27	0.45	-1.08	0.19	-0.13	0.0015
275	1.31	1.43	0.12	**290**	1.17	0.92	-0.55	0.20	-0.12	0.0014
276	1.29	0.81	0.17	**291**	1.15	0.59	0.46	0.19	-0.13	0.0014
277	1.12	1.31	0.14	**292**	1.12	0.18	0.10	0.19	-0.12	0.0011
278	1.02	0.88	0.17	**293**	1.08	0.50	-0.05	0.16	-0.12	0.0011
279	1.01	-0.68	0.15	**294**	1.05	-0.68	-0.33	0.22	-0.11	0.0010
280	1.00	-0.45	0.14	**295**	1.04	0.07	-0.26	0.15	-0.11	0.0010
281	0.98	0.78	0.14	**296**	0.99	0.35	0.52	0.21	-0.11	0.0010
282	0.94	-0.05	0.17	**297**	0.96	-0.15	-0.22	0.15	-0.11	0.0009
283	0.91	-0.20	0.12	**298**	0.96	0.04	0.05	0.15	-0.12	0.0009
284	0.88	-0.86	0.18	**299**	0.91	-0.60	-1.00	0.18	-0.11	0.0009
285	0.85	-0.84	0.17	**300**	0.91	0.14	-0.36	0.15	-0.11	0.0009
Media	*1.14*	*0.35*	*0.16*		*1.13*	*0.17*	*-0.23*	*0.17*	*-0.12*	0.0012
Desv.std.	*(0.24)*	*(0.86)*	*(0.02)*		*(0.20)*	*(0.47)*	*(0.47)*	*(0.03)*	*(0.01)*	*(0.0004)*

6.5.4. Factores manipulados

Contaminación del banco: En la condición sin DIF (Banco no contaminado), los parámetros *b* del grupo de referencia y focal coincidían para los 30 ítems (del 271 al 300). En la condición con DIF (Banco contaminado), en el 5% de los ítems (del 286 al 300) el parámetro *b* difería entre grupos (ver Tabla 6.1). En esta condición, la selección de ítems y la estimación del nivel de rasgo para ambos grupos se hacían con los parámetros del grupo de referencia. Las probabilidades de acierto para simular las respuestas a los ítems se calculaban con los parámetros según el grupo de pertenencia.

- Presencia de impacto: Se simularon dos condiciones, una con impacto y otra sin impacto. En ambas, la distribución del nivel de habilidad de los examinados del grupo de referencia seguía una distribución normal estándar. Para el grupo focal, en la condición sin impacto, la distribución era igual a la del grupo de referencia, mientras que en la condición con impacto era $N(-0.5, 1)$.
- Tamaño de las muestras: El número total de examinados incluidos en el análisis se fijó en 1,000. En una primera condición, el número de examinados por grupo fue igual. En la otra condición, N_R fue igual a 750 y N_F fue igual a 250. Para ello, se seleccionaban, de entre los 10,000 examinados, los primeros que habían respondido al ítem a analizar, hasta completar el tamaño requerido. Por tanto, la submuestra empleada cambia de ítem a ítem.

Se generaron 100 réplicas para cada una de las 8 condiciones resultantes de combinar presencia de DIF con presencia de impacto y tamaño de las muestras. Cada una de estas réplicas fue analizada con los tres métodos de detección de DIF: IRT LRT, CATSIB con corrección (CATSIB-C) y CATSIB sin corrección (CATSIB-SC). En todos los casos, el valor de α se fijó en 0.05.

- IRT LRT: El software de calibración para los modelos en IRT fue el ICL (Hanson, 2002). Se construyó el modelo compacto y el aumentado para el análisis de cada ítem. En ambos modelos, se fijaban los valores de los parámetros de los ítems no analizados a sus valores conocidos; en el modelo compacto se restringían los parámetros del ítem estudiado a ser iguales entre los dos grupos, mientras que en el modelo aumentado se permitía que fueran distintos. Se implementó el método de estimación en el que se actualiza

múltiples veces la distribución previa y en el que se utilizan múltiples ciclos EM. Se definieron distribuciones previas para los parámetros *a* y *c*: para el *a*, lognormal(0, 0.5); para el *c*, beta(4, 16, 0, 1), donde los primeros dos valores son los parámetros para la forma y los dos siguientes el límite inferior y superior. Estas distribuciones son comunes a otros programas de estimación como PARSCALE o BILOG (du Toit, 2003). Se estableció un máximo de 1,000 iteraciones EM para los procedimientos con MEM y un criterio de convergencia de 0.001.

- CATSIB: Se siguió la descripción de Nandakumar y Roussos (2001, 2004). Para la comparación entre grupos se definieron 80 intervalos iniciales que dividían el continuo de habilidad estimada, $\hat{\theta}$ para CATSIB-SC o $\hat{\theta}*$ para CATSIB-C. Los intervalos se consideraban para el cálculo de $\hat{\beta}_{CATSIB}$ si tenían al menos tres evaluados por grupo; en caso contrario, se descartaban. Si el porcentaje de eliminados en un grupo era mayor que 7.5%, se reducía el número de intervalos hasta que el porcentaje de intervalos descartados era menor que 7.5% para ambos grupos o hasta un mínimo de 20.

6.5.5. Criterios de valoración

- Detección del DIF: La calidad de la detección del DIF fue evaluada mediante el error Tipo I y la potencia. El error Tipo I (falsos positivos), fue valorado en la condición sin DIF para los 30 ítems analizados y en el banco contaminadon para los 15 ítems analizados libres de DIF. La condición con DIF permite valorar el efecto de la contaminación del test de anclaje. El IC 95% del valor de α, dado el número de replicas, es [0.007, 0.093]. La potencia (detección de DIF habiéndolo) se consideró como aceptable si era igual o superior a 0.80.

- Recuperación del tamaño del DIF:

- El tamaño real se calculó como:

$$\beta_j = \sum_{k=1}^{K} (P_j(u_j = 1 \mid \theta_k, \boldsymbol{\Delta}_{jR}, g = R) - P_j(u_j = 1 \mid \theta_k, \boldsymbol{\Delta}_{jF}, g = F)) f_j(\theta_k \mid \boldsymbol{\pi}_{(j)}) \qquad (6.9)$$

donde $\boldsymbol{\pi}_{(j)}$ es el vector con la media y la desviación típica en θ del grupo total que ha respondido al ítem j ($\boldsymbol{\mu}_{(j)}, \boldsymbol{\sigma}_{(j)}$) y $f_j(\theta_k \mid \boldsymbol{\pi}_{(j)})$ es la función de distribución normal discretizada. Se utilizaron 80 puntos de cuadratura (K = 80).

El valor estimado se obtuvo por cada uno de los tres métodos. El tamaño del efecto calculado en el IRT LRT fue:

$$\hat{\beta}_{j(IRTLR)|TAI} = \sum_{k=1}^{K} (P_j(u_j = 1 \mid \theta_k, \hat{\boldsymbol{\Delta}}_{jR}, g = R) - P_j(u_j = 1 \mid \theta_k, \hat{\boldsymbol{\Delta}}_{jF}, g = F)) f_j(\theta_k \mid \hat{\boldsymbol{\pi}}_{(j)}) \qquad (6.10)$$

donde $\hat{\boldsymbol{\Delta}}_{jR}$ y $\hat{\boldsymbol{\Delta}}_{jF}$ son los parámetros de los ítems estimados en el modelo aumentado y $\hat{\boldsymbol{\pi}}_{(j)}$ es el vector con la media y la desviación típica en θ del grupo total que ha respondido al ítem j ($\hat{\boldsymbol{\mu}}_{(j)}, \hat{\boldsymbol{\sigma}}_{(j)}$), calculadas a partir de los parámetros de distribución obtenidos en el modelo aumentado.

Los tamaños del efecto estimados en CATSIB-C ($\hat{\beta}_{j(CATSIB-C)}$) y CATSIB-SC ($\hat{\beta}_{j(CATSIB-SC)}$) fueron estimados según las ecuaciones X y X, estableciendo el número de intervalos siguiendo los criterios ya descritos (un mínimo de 20).

La calidad de la recuperación de β se evaluó mediante el sesgo y el RMSE.

6.5.6. Análisis

Se realizaron ANOVAs para determinar qué factores afectaban a la tasa de error Tipo I, a la potencia y a la recuperación del tamaño del DIF. Se utilizó un ANOVA mixto en el que los procedimientos de estimación empleados se consideran una medida repetida, mientras que las variables tamaño de la muestra e impacto se consideran medidas independientes. La variable dependiente era la proporción de veces que se detectaba DIF a través de las réplicas (para el análisis de las tasas de error tipo I y de potencia) o el sesgo y el RMSE promedio a través de las réplicas en la recuperación de $\tilde{\beta}$

Todos los efectos informados son significativos con $p < 0.05$. Como medida de tamaño del efecto se utilizó η^2 parcial. Únicamente se informa de aquellos efectos con tamaños del efecto superiores a 0.14 que, en la clasificación de Cohen (1988, 1992), son considerados grandes.

6.6. Resultados

6.6.1. Tasas de error Tipo I en la condición sin DIF (Banco no contaminado)

En las Tablas 6.2 y 6.3 se presentan las tasas de error para cada técnica de detección de DIF en los ítems del 271 al 285 y para los ítems del 286 al 300 según tamaño de las muestras y presencia de impacto. Las celdas sombreadas corresponden a valores fuera del intervalo de confianza del nivel nominal permitido. IRT LRT mostró un adecuado control de la tasa de error en todos los ítems analizados, si bien la prueba resultó excesivamente conservadora, con tasas de error Tipo I marcadamente por debajo del valor nominal. Cuando el DIF se analiza con CATSIB algunos ítems presentan tasas de error por encima del límite del IC, pero la tasa promedio de error Tipo I se ajusta en gran medida al valor nominal. El ANOVA mostró efectos significativos relevantes del método [$F(2,232) = 171.96$, $p < 0.001$; η^2=.60], siendo menor el error tipo I de IRT LRT ($\bar{x}_{IRT-LRT} = 0.023$; $\bar{x}_{CATSIB-C} = \bar{x}_{CATSIB-SC} = 0.060$). También se encontró un mayor efecto del impacto en la inflación de las tasas de error tipo I al aplicar el método CATSIB sin corrección en comparación al IRT LRT [$F(1,116) = 18.64$, $p < 0.001$; η^2=0.14; para IRT LRT: $\bar{x}_{\text{sinimpacto}}$=0.022, $\bar{x}_{\text{conimpacto}}$=0.023; para CATSIB-SC: $\bar{x}_{\text{sinimpacto}}$=0.049, $\bar{x}_{\text{conimpacto}}$=0.070).

Se puso a prueba si los resultados del ANOVA dependían de los grupos de ítems formados (del 271-285 y del 286-300) introduciendo esta nueva variable, sin encontrar un efecto significativo [$F(2,230) = 2.14$, $p = 0.12$].

Tabla 6.2. Tasas de error Tipo I para los conjuntos de datos libres de DIF. Ítems del 271 al 285

	Métodos											
	IRT LRT				CATSIB CON CORRECCIÓN				CATSIB SIN CORRECCION			
	Sin impacto		Impacto		Sin impacto		Impacto		Sin impacto		Impacto	
ITEM	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500
271	0.01	0.03	0.02	0.03	0.04	0.02	0.06	0.05	0.03	0.02	0.01	0.06
272	0.01	0.01	0.00	0.02	0.06	0.05	0.08	0.07	0.05	0.04	0.08	0.09
273	0.03	0.01	0.05	0.04	0.04	0.04	0.05	0.04	0.04	0.03	0.07	0.07
274	0.01	0.04	0.05	0.02	0.05	0.04	0.07	0.08	0.04	0.05	0.06	0.08
275	0.02	0.03	0.01	0.03	0.06	0.01	0.12	0.04	0.05	0.01	0.10	0.07
276	0.01	0.03	0.01	0.04	0.06	0.04	0.04	0.03	0.06	0.05	0.07	0.06
277	0.04	0.03	0.04	0.02	0.05	0.02	0.04	0.02	0.08	0.02	0.10	0.06
278	0.00	0.01	0.06	0.04	0.06	0.04	0.10	0.11	0.06	0.03	0.07	0.13
279	0.02	0.05	0.00	0.01	0.04	0.10	0.01	0.06	0.06	0.08	0.04	0.07
280	0.01	0.01	0.03	0.01	0.05	0.03	0.06	0.06	0.06	0.01	0.06	0.04
281	0.01	0.03	0.01	0.04	0.05	0.05	0.05	0.07	0.05	0.07	0.06	0.08
282	0.03	0.03	0.01	0.03	0.03	0.04	0.04	0.06	0.05	0.03	0.08	0.13
283	0.00	0.02	0.02	0.01	0.03	0.02	0.05	0.01	0.01	0.02	0.05	0.06
284	0.01	0.03	0.02	0.03	0.06	0.05	0.06	0.07	0.04	0.07	0.05	0.06
285	0.01	0.04	0.02	0.04	0.04	0.11	0.13	0.08	0.04	0.10	0.05	0.08
Media	*0.01*	*0.03*	*0.02*	*0.03*	0.05	0.04	0.06	0.06	0.05	0.04	0.06	0.08
Desv.std.	*(0.011)*	*(0.012)*	*(0.018)*	*(0.012)*	*(0.011)*	*(0.028)*	*(0.032)*	*(0.026)*	*(0.016)*	*(0.027)*	*(0.023)*	*(0.025)*

Nota: Los ítems que no tuvieron un adecuado control de la tasa de error tipo I en las condiciones simuladas se muestran sombreadas

Tabla 6.3. Tasas de error Tipo I para los conjuntos de datos libres de DIF (banco no contaminado). Ítems del 286 al 300

	Métodos											
	IRT LRT				CATSIB CON CORRECCIÓN				CATSIB SIN CORRECCION			
	Sin impacto		Impacto		Sin impacto		Impacto		Sin impacto		Impacto	
Item	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500
286	0.05	0.04	0.02	0.02	0.09	0.05	0.14	0.06	0.07	0.06	0.05	0.07
287	0.04	0.03	0.02	0.05	0.07	0.05	0.03	0.04	0.06	0.05	0.05	0.04
288	0.01	0.01	0.01	0.01	0.07	0.05	0.03	0.05	0.07	0.05	0.06	0.05
289	0.02	0.06	0.02	0.04	0.05	0.06	0.10	0.12	0.03	0.07	0.10	0.09
290	0.04	0.00	0.02	0.03	0.01	0.08	0.04	0.06	0.02	0.06	0.04	0.08
291	0.02	0.04	0.01	0.03	0.10	0.07	0.06	0.06	0.10	0.07	0.07	0.05
292	0.00	0.03	0.01	0.03	0.02	0.09	0.04	0.06	0.01	0.08	0.06	0.08
293	0.04	0.01	0.03	0.02	0.09	0.05	0.06	0.04	0.09	0.05	0.05	0.06
294	0.00	0.01	0.02	0.01	0.06	0.07	0.08	0.02	0.05	0.04	0.09	0.10
295	0.02	0.00	0.00	0.04	0.05	0.07	0.04	0.06	0.05	0.03	0.05	0.04
296	0.02	0.01	0.00	0.02	0.05	0.08	0.09	0.08	0.05	0.09	0.09	0.07
297	0.04	0.01	0.03	0.01	0.06	0.04	0.04	0.05	0.06	0.03	0.06	0.05
298	0.03	0.03	0.02	0.01	0.05	0.05	0.09	0.05	0.05	0.05	0.11	0.07
299	0.02	0.02	0.03	0.02	0.03	0.03	0.08	0.10	0.00	0.07	0.10	0.09
300	0.03	0.02	0.04	0.03	0.08	0.07	0.08	0.07	0.06	0.05	0.07	0.11
Media	*0.03*	*0.02*	*0.02*	*0.02*	*0.06*	*0.06*	*0.07*	*0.06*	*0.05*	*0.06*	*0.07*	*0.07*
Desv.std.	*(0.015)*	*(0.016)*	*(0.011)*	*(0.012)*	*(0.025)*	*(0.016)*	*(0.031)*	*(0.024)*	*(0.027)*	*(0.017)*	*(0.024)*	*(0.021)*

Nota: Los ítems que no tuvieron un adecuado control de la tasa de error tipo I en las condiciones simuladas se muestran sombreada

6.6.2. Tasas de error tipo I en las condiciones con DIF (Banco contaminado)

En la Tabla 6.4 se presentan las tasas de error Tipo I para los ítems del 271 al 285 en la condición en la que el otro grupo de ítems presenta DIF. En términos generales, la contaminación de ítems con DIF en el banco incrementa las tasas de error tipo I para todos los métodos, sobre todo para CATSIB, método que pasa a mostrar un porcentaje considerablemente alto de ítems con tasas de error mayores al nivel nominal permitido. Por el contrario, IRT LRT sigue manteniendo, a excepción de un ítem en una condición, un control adecuado de las tasas de error. El ANOVA indicó un efectos relevantes para el método con el que se analizó el DIF [$F(2,112) = 237.77$, $\eta_p^2 = 0.81$] y para la interacción entre el método y el impacto [$F(2,112) = 61.37$, $\eta_p^2 = 0.52$]. En general, se encontró que las tasas de error Tipo I obtenidas con IRT LRT eran significativamente menores a las obtenidas en CATSIB-C [$F(1,56) = 251.98$, $\eta_p^2 = 0.82$] o en CATSIB-SC [$F(1,56) = 332.13$, $\eta_p^2 = 0.86$]. La naturaleza de la interacción puede observarse en la Figura 6.4, en la que se presentan las tasas de error Tipo I en función del impacto para cada uno de los métodos. Se observa que la presencia de impacto incrementa considerablemente las tasas de error Tipo I para CATSIB cuando no se incluye la corrección ($\bar{x}_{\text{sinimpacto}}$=0.10, $\bar{x}_{\text{conimpacto}}$=0.16), mientras que esta variable apenas influye para el IRT LRT ($\bar{x}_{\text{sinimpacto}}$=0.04, $\bar{x}_{\text{conimpacto}}$=0.03) o para el CATSIB con corrección ($\bar{x}_{\text{sinimpacto}}$=0.11, $\bar{x}_{\text{conimpacto}}$=0.09).

Figura 6.1. Tasa de error Tipo I local según método de detección de DIF e impacto.

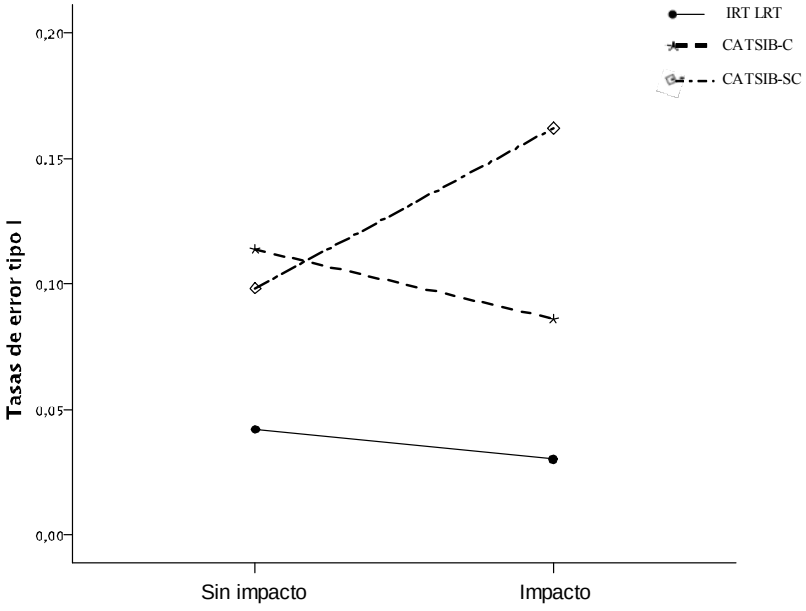


Tabla 6.4. Tasas de error Tipo I para los conjuntos de datos con DIF. Ítems del 271 al 285.

	Métodos											
	IRT LRT				CATSIB CON CORRECCIÓN				CATSIB SIN CORRECCION			
	Sin impacto		Impacto		Sin impacto		Impacto		Sin impacto		Impacto	
ITEM	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500
271	.03	.09	.02	.11	.18	.19	.03	.21	.12	.15	.11	.31
272	.02	.04	.01	.03	.08	.10	.10	.05	.06	.11	.17	.15
273	.03	.09	.09	.05	.12	.14	.13	.11	.13	.14	.19	.28
274	.03	.02	.01	.04	.11	.12	.05	.08	.13	.12	.12	.14
275	.03	.08	.03	.01	.13	.10	.05	.12	.12	.08	.16	.21
276	.05	.06	.04	.02	.13	.14	.09	.15	.12	.12	.22	.25
277	.01	.04	.01	.02	.06	.11	.08	.08	.06	.11	.17	.19
278	.04	.05	.00	.04	.11	.12	.07	.10	.11	.11	.10	.20
279	.02	.04	.02	.03	.10	.13	.05	.02	.07	.12	.13	.07
280	.01	.01	.04	.00	.10	.05	.08	.04	.07	.03	.18	.16
281	.03	.06	.02	.03	.10	.14	.07	.11	.07	.13	.14	.15
282	.05	.04	.07	.04	.12	.10	.12	.13	.10	.09	.15	.23
283	.02	.05	.02	.04	.12	.19	.09	.10	.09	.13	.14	.19
284	.05	.06	.00	.01	.09	.07	.08	.04	.08	.07	.05	.06
285	.06	.05	.02	.03	.08	.08	.07	.08	.07	.04	.08	.16
Media	*.03*	*.05*	*.03*	*.03*	*.11*	*.12*	*.08*	*.09*	*.09*	*.10*	*.14*	*.18*
Desv.std.	*.011*	*.012*	*.018*	*.012*	*.011*	*.028*	*.032*	*.026*	*.016*	*.027*	*.023*	*.025*

Nota: Los ítems que no tuvieron un adecuado control de la tasa de error tipo I en las condiciones simuladas se muestran sombreadas.

Tabla 6.5. Tasas de Potencia para los ítems con DIF.

	Métodos											
	IRT LRT				CATSIB CON CORRECCIÓN				CATSIB SIN CORRECCION			
	Sin impacto		Impacto		Sin impacto		Impacto		Sin impacto		Impacto	
Item	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500	750-250	500-500
286	**0.90**	**0.98**	**0.95**	**0.98**	0.96	0.99	0.96	1.00	0.97	0.99	0.95	0.98
287	**0.82**	**0.98**	**0.83**	**0.98**	0.94	1.00	0.96	1.00	0.93	1.00	0.93	0.97
288	0.79	**0.94**	0.77	**0.95**	0.91	0.99	0.90	0.99	0.93	0.99	0.83	0.96
289	0.74	**0.85**	**0.84**	**0.93**	0.95	0.98	0.95	0.99	0.95	0.98	0.88	0.96
290	0.73	**0.87**	0.79	**0.86**	0.89	0.94	0.93	0.97	0.89	0.93	0.88	0.91
291	0.71	**0.86**	0.71	**0.87**	0.89	0.91	0.90	0.96	0.89	0.94	0.81	0.93
292	0.65	**0.90**	0.67	**0.85**	0.88	0.95	0.92	0.95	0.90	0.97	0.84	0.90
293	0.68	**0.84**	0.74	**0.85**	0.91	0.93	0.95	0.96	0.93	0.95	0.85	0.89
294	0.72	**0.85**	0.76	**0.87**	0.93	0.95	0.90	0.97	0.92	0.95	0.81	0.95
295	0.55	0.68	0.61	0.75	0.79	0.89	0.88	0.93	0.79	0.90	0.78	0.86
296	0.54	0.79	0.52	0.79	0.79	0.92	0.88	0.95	0.80	0.96	0.77	0.89
297	0.59	0.69	0.71	**0.82**	0.83	0.89	0.86	0.96	0.84	0.90	0.78	0.90
298	0.59	0.70	0.55	0.77	0.82	0.89	0.73	0.89	0.87	0.91	0.64	0.85
299	0.67	0.73	0.71	**0.85**	0.86	0.92	0.90	0.95	0.87	0.92	0.76	0.87
300	0.46	0.60	0.56	0.68	0.79	0.88	0.81	0.92	0.80	0.88	0.69	0.81
Media	*0.68*	*0.82*	*0.71*	*0.85*	*0.88*	*0.94*	*0.90*	*0.96*	*0.89*	*0.94*	*0.81*	*0.91*
Desv.std.	*(0.015)*	*(0.016)*	*(0.011)*	*(0.012)*	*(0.025)*	*(0.016)*	*(0.031)*	*(0.024)*	*(0.027)*	*(0.017)*	*(0.024)*	*(0.021)*

Nota: Los métodos que no tuvieron un adecuado control de la tasa de error tipo I en las condiciones simuladas se muestran sombreadas. En IRT LRT las tasas de potencia ≥ 0.80 están señalas en negritas.

6.6.3. Potencia

En la Tabla 6.5 se presentan las tasas de potencia de detección de DIF para cada uno de los métodos según impacto y tamaño muestral. La descripción de los resultados de potencia se analizan considerando el comportamiento del error Tipo I. Dado los problemas del CATSIB en el control de las tasas de error Tipo I señaladas previamente, las tasas de potencia calculadas para éste método no se analizan (valores de potencia sombreados).

En la Tabla 6.4 las potencias para IRT LRT mayores a .80 se destacan con negrita. El ANOVA reveló un efecto significativo sobre la potencia en IRT LRT para el tamaño de las muestras [$F(1,56) = 24.37$, $\eta_p^2 = 0.30$]. Específicamente se encontró que la potencia era significativamente mayor cuando el tamaño de las muestras era igual ($\bar{x}_{500\text{-}500} = 0.84$) que cuando eran diferentes ($\bar{x}_{750\text{-}250} = 0.70$). De manera sistemática las tasas de potencia mayores se registraron en los ítems más discriminativos.

6.6.4. Recuperación del tamaño del DIF

En la Tabla 6.6 se presentan los valores del sesgo y del RMSE promedio para el tamaño del DIF para cada uno de los métodos en las condiciones de impacto y tamaños de muestra. En el análisis para el sesgo se encontró un efecto para el método [$F(2,232) = 222.16$, $\eta_p^2 = 0.66$] y para la interacción entre el método y el impacto [$F(2,232) = 344.93$, $\eta_p^2 = 0.75$].

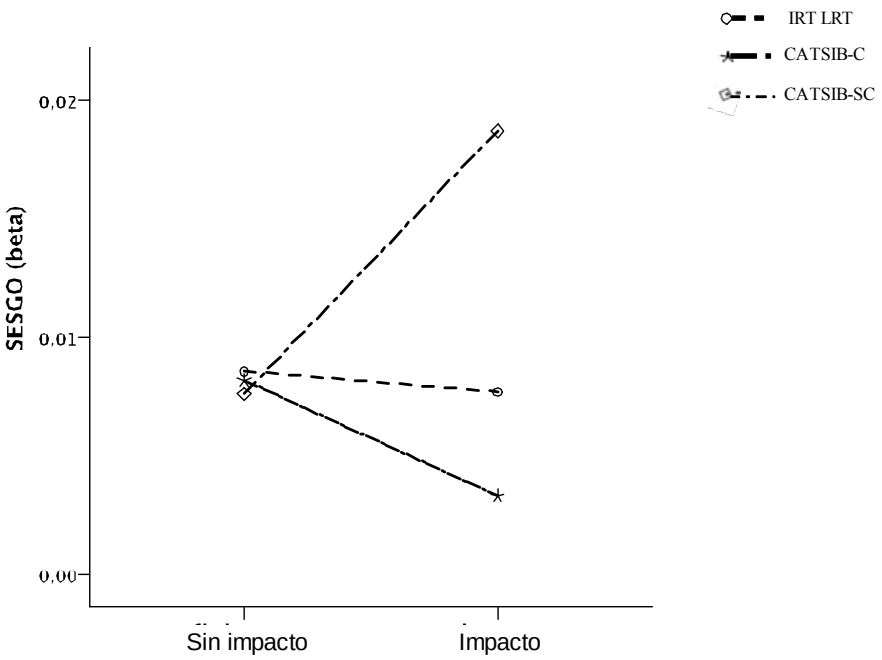
Tabla 6.6. Sesgo y RMSE promedio para el tamaño del DIF por impacto y tamaño de las muestras en cada método.

		METODOS		
	Muestras	IRTLRT	CATSIB-C	CATSIB-SC
			Sesgo	
Sin impacto	750-250	0.0168	0.0160	0.0148
		(0.0054)	(0.0059)	(0.0051)
	500-500	0.0187	0.0173	0.0158
		(0.0054)	(0.0057)	(0.0052)
Impacto	750-250	0.0176	0.0121	0.0267
		(0.0066)	(0.0088)	(0.0069)
	500-500	0.0173	0.0125	0.0267
		(0.0061)	(0.0077)	(0.0061)
			RMSE	
Sin impacto	750-250	0.0365	0.0387	0.0377
		(0.0026)	(0.0031)	(0.0030)
	500-500	0.0345	0.0353	0.0342
		(0.0029)	(0.0032)	(0.0031)
Impacto	750-250	0.0368	0.0373	0.0434
		(0.0040)	(0.0042)	(0.0052)
	500-500	0.0333	0.0330	0.0400
		(0.0038)	(0.0045)	(0.0051)

Nota: Los valores entre paréntesis corresponden a la desviación típica

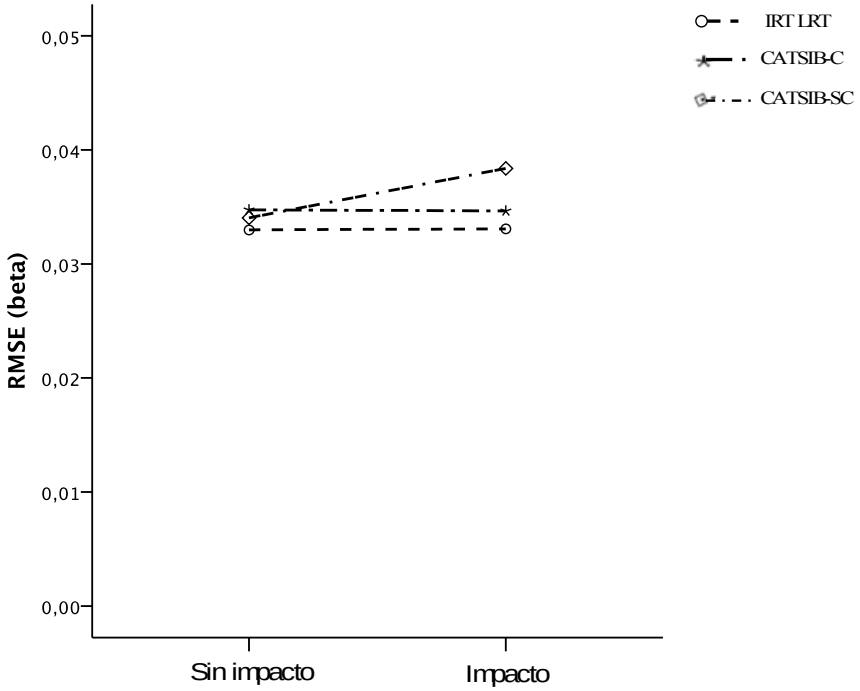
En la Figura 6.2 se presenta el sesgo como función del impacto. Como puede observarse, en ausencia de impacto no hay diferencias entre los métodos ($\bar{x}_{IRT-LRT} = .018; \bar{x}_{CATSIB-C} = .017; \bar{x}_{CATSIB-SC} = .015$). Sin embargo, el sesgo se incrementa significativamente para CATSIB sin corrección en la condición de impacto y disminuye para CATSIB con corrección ($\bar{x}_{IRT-LRT} = .017; \bar{x}_{CATSIB-C} = .012; \bar{x}_{CATSIB-SC} = .027$).

Figura 6.2. Sesgo en la estimación de β en la condición con DIF, como función del impacto



De la misma forma el ANOVA aplicado al RMSE mostró un efecto para el método con el que calculó el tamaño del DIF [$F(2,232) = 192.67$, $\eta_p^2 = 0.62$], así como una interacción entre el método y el impacto [$F(2,232) = 229.86$, $\eta_p^2 = 0.67$]. En la Figura 6.6 se presenta el RMSE en función del impacto para cada uno de los métodos. En ausencia de impacto no se observan diferencias significativas ($\bar{x}_{IRT-LRT} = .036; \bar{x}_{CATSIB-C} = .037; \bar{x}_{CATSIB-SC} = .036$) entre los métodos, mientras que CATSIB sin corrección muestra un error de estimación mayor para la condición de impacto ($\bar{x}_{IRT-LRT} = .035; \bar{x}_{CATSIB-C} = .035; \bar{x}_{CATSIB-SC} = .042$)

Figura 6.3. RMSE en la estimación de β en la condición con DIF, como función del impacto



6.7. Discusión y conclusiones

En el presente trabajo se analizó la viabilidad de utilizar IRT LRT y CATSIB (con corrección y sin corrección) para el estudio del DIF en ítems aplicados adaptativamente. Se simuló un escenario en el que un grupo de ítems con mayor tasa de exposición presentaba una dificultad menor en el grupo focal respecto del grupo de referencia, algo que podría darse en el caso de diseminación del contenido de parte del banco de ítems. Tres factores se manipularon para comparar la efectividad de los métodos: la presencia de DIF en otros ítems del banco (contaminación), el impacto y el tamaño de las muestras. La efectividad de los métodos se evalúo en términos de la identificación correcta de los ítems con DIF y la recuperación del tamaño del DIF.

La identificación correcta del DIF en los ítems del banco, especialmente de aquellos ítems con altas tasas de exposición es fundamental en los estudios de validez. Se sabe que el algoritmo adaptativo suele implicar un uso considerable de los ítems más discriminativos (Li y Schafer, 2005). La eliminación de los ítems más informativos se

convierte en una decisión difícil en el ámbito aplicado porque la renovación es un proceso costoso, especialmente si se desea mantener la capacidad discriminativa del banco de ítems (Hau y Chang, 2001). Por lo tanto, es importante que los métodos que se aplican al análisis del DIF controlen el grado en que se realizan detecciones incorrectas en al menos tres situaciones: (a) en ausencia de DIF; (b) cuando algunos ítems del banco presentan DIF; y (c) en presencia de impacto.

En la situación con ausencia de DIF, IRT LRT tuvo mayor capacidad para mantener un control de la tasa de error Tipo I en todos los ítems independientemente del impacto y del tamaño de las muestras. Estos resultados replican lo obtenido por Lei et al. (2006) en su estudio sobre la detección del DIF en ítems pretest. CATSIB tuvo un comportamiento razonable, pero más irregular que se observa en el rango de tasas de error obtenidas, entre 0.01 y 0.13. Los resultados de nuestro estudio difieren de los obtenidos por Lei et al. (2006) que encontraron tasas de error Tipo I más altas en las condiciones con tamaños muestrales distintos, especialmente en presencia de impacto. En nuestro estudio, 20 intervalos resultaron adecuados incluso con tamaños muestrales desiguales. Las diferencias con el estudio de Lei et al. (2006) pueden deberse a que la aplicación adaptativa genera, incluso en presencia de impacto, mayor solapamiento de las distribuciones del grupo focal y de referencia que responde al ítem. Otra posible explicación es que las condiciones en su estudio eran más extremas (muestras de 100 sujetos en el grupo focal, ítems con niveles de dificultad muy extremos y un mayor nivel de impacto).

Como se ha señalado, los métodos deben demostrar no solo un control de la tasa de error Tipo I cuando el banco está libre de DIF, sino también cuando existe contaminación en el banco. En esta condición, IRT LRT mantuvo un control de las tasas de error tipo I adecuado, similar al encontrado en la condición sin DIF, mientras que CATSIB presenta una pérdida de control en las tasas de error tipo I en la mayoría de los ítems. El incremento del error Tipo I es un resultado esperable, en presencia de contaminación en el TAI, para cualquier técnica de DIF. En la investigación en tests fijos, estos efectos son comunes salvo cuando son pocos los ítems con DIF o cuando el funcionamiento diferencial del test implica que la prueba no favorece a ninguno de los grupos (Bolt y Gierl, 2006; Gierl, Gotzmann, Boughton, 2004, Shealy y Stout, 1993b; Roussos y Stout, 1996). En este estudio, los ítems con DIF están entre los más expuestos y el DIF es siempre en el mismo sentido. Por tanto, la presencia de ítems de

anclaje con DIF sesga en cierto grado la estimación de $\hat{\theta}_{TAI}$ en CATSIB en el grupo focal.

Por el contrario, el excelente resultado para IRT LRT cuando hay contaminación del banco es contrario a lo esperado. En estudios previos, en los métodos de detección de DIF basados en modelos (como el IRT LRT) la contaminación produce imprecisiones en la estimación de los parámetros de los ítems y en la distribución de la habilidad de los grupos, lo que conduce a una sobreestimación del tamaño del DIF, causando tasas de error Tipo I extremas (Finch, 2005; Stark, Chernyshenko y Drasgow, 2006; Wang, 2004, Wang y Yeh, 2003, Woods, 2008a). Una posible explicación al hecho de que la contaminación no afecte a las tasas de error Tipo I en IRT LRT es el procedimiento de calibración utilizado. En la CPF los parámetros de los ítems del test de anclaje no se estiman, sino que se fijan a los valores previamente estimados, lo que implica que, al menos, los parámetros de los ítems de anclaje sin DIF no se ven comprometidos. Estos suponen una parte importante del banco (el 95%), lo que podría tener un efecto positivo en el control de la tasa de falsos positivos.

Otro de las problemas importantes en la aplicación de los métodos clásicos es conseguir definir una variable de igualación apropiada, que evite la inflación de las tasas de error Tipo I en presencia de impacto (Zwick, 2001, 2007). La inclusión o no del ítem analizado para el estudio del DIF en ítems pretest (Zwick, Thayer y Wingersky, 1995) o la exclusión de la corrección por regresión en el análisis del DIF en ítems operativos, así como el uso de procedimientos alternativos, se considera un campo en revisión para este tipo de métodos (Monahan y Ankenmann, 2010; Walker et al., 2001). Nuestro estudio muestra que la inclusión del ítem estudiado en la estimación del rasgo no es suficiente para corregir la variable de igualación y evitar tasas de error Tipo I altas, lo que fue más evidente ante la presencia de DIF en el banco. En presencia de DIF e impacto, las tasas de error Tipo I, fueron considerablemente más altas cuando no se utilizó la corrección. De la misma forma el sesgo y el RMSE en la recuperación del tamaño del DIF aumentaron en esa condición.

Respecto a la potencia, la falta de control en las tasas de error Tipo I para CATSIB hace poco interpretables los resultados sobre detección de DIF para este método. Por el contrario, IRT LRT muestra una potencia adecuada en la detección del DIF de ítems aplicados adaptativamente, especialmente cuando el tamaño de las

muestras fue de 500 en cada grupo. En este caso, las tasas de potencia fueron aceptables para un porcentaje amplio de los ítems estudiados.

En relación con la recuperación del tamaño del DIF se encuentra que IRT LRT obtiene mejores resultados a los encontrados con CATSIB. Este resultado puede deberse, parcialmente, al ajuste de los datos al modelo paramétrico para la generación de datos.

Teniendo en cuenta el patrón de resultados, nuestra recomendación es utilizar IRT LRT para el análisis de DIF en ítems operativos, puesto que: (a) proporciona un control adecuado de la tasa de error Tipo I , independientemente de la condición simulada; (b) la potencia es adecuada para tamaños muestrales razonables; y (c) encontramos una buena recuperación del tamaño del DIF. Ciertamente, las sugerencias y conclusiones están limitadas a estas condiciones de simulación. En el presente estudio se simularon condiciones realistas en términos de los parámetros del banco de ítems, de la longitud del TAI en relación con el tamaño del banco e incluyendo en el algoritmo adaptativo, métodos de selección de ítems y de control de la exposición que favorecieran la seguridad del test. Al mismo tiempo los tamaños de muestra y el impacto son bastante plausibles en los estudios aplicados. Pero se considera que es necesario tener mayor evidencia que permita conocer el comportamiento de los métodos en el análisis de ítems con otras características (parámetros distintos), estudiar tipos de DIF distintos, así como situaciones de contaminación que se adecuen a lo esperado en TAIs.

6.8. Referencias

Abad, F., Olea, J., Aguado, D., Ponsoda, V., Barrada, J. (2010). Deterioro de parámetros de los ítems en test adaptativos informatizados: estudio con eCAT. *Psicothema*. *22*, 340-347.

Bolt, D., y Gierl, M. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement, 43*, 313-334.

Camilli, G., y Shepard, L. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Cheng, Y. (2009). Computerized adaptive testing for cognitive diagnosis. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieval [13-04-2011] from www.psych.umn.edu/psylabs/CATCentral/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* 2ª ed. Hillsdale, NJ. Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cuevas, L. Abad, J. Olea, J., Barrada, J. y Garrido, L.(2010). *CATSIM 1.0.* Software presented at the IV European Congress of Methodology. Potsdam, Germany.

DeMars, Ch. (2004). Detection of item parameter Drift over multiple test administrations. *Applied Measurement in Education. 17*, 265-300.

du Toit, M. (Ed.), (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Embretson, S. (2004). The second century of ability testing: some predictions and speculations. *Measurement, 2,* 1-32.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295.

Gierl, M., Gotzmann, A., Boughton, K. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17*, 241-264.

Glass, C. (2007). Item parameter estimation and item fit analysis. En W. J. van der Linden y C. A. W. Glas (Ed.), *Computerized Adaptive Testing. Theory and Practice* (págs. 269-288). Boston, MA: Kluwer Academic Publishers.

Hanson, B. (2002). *IRT Command Language* (Version 0.020301). Monterey, CA: Author. (Available at http://sourceforge.net/projects/ssm)

Harmes, J., Kromrey, J. y Parshall, C.(2001). *Online item parameter recalibration: application of missing data treatments to overcome the effects of sparse data conditions in a computerized adaptive version of the MCAT*. Report submitted to the Association of American Medical Colleges. Section for the MCAT. 1-27.

Hart, D., Deutsher, D., Crane, P. y Wang, Y. (2009). Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairment who spoke English or Hebrew. *Quality of Life Research. 18*, 1067-1083.

Hau, K., y Chang, H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266.

Haynie, K.A. y Way, W.D. (1995). *An investigation of item calibration procedure for a computerized licensure examination.* Paper presented at a symposium entitled Computer Adaptive Testing, at the annual meeting of the National Council on Measurement in Education, San Francisco.

Holland, P.y Thayer, D. (1988). Differential item performance and the Mantel–Haenszel statistic. In: H. Wainer y H. Braun (Ed.), *Test Validity*, Hillsdale: Erlbaum, pp. 129–145.

Ito, K. y Sykes, R.C. (1994). *The effect of restricting ability distributions int he estimation of item difficulties: Implications for a CAT implementation.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Jodoin, M.G., y Gierl, M.J. (2001). Evaluating Type 1 error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*, 355-381.

Kingsbury, G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieval [13-04-2011] from www.psych.umn.edu/psylabs/CATCentral/

Kingsbury, G.G., y Houser, R.L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice*, *12,* 21-27.

Lei, P.-W., Chen, S.-Y. y Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43,* 245–264.

Li, Y., y Schafer, W. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42*, 245-269.

Mills, C. y Stocking, M. (1996). Practical Issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*, 287-304

Nandakumar, R. y Roussos, L. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests* (Research report). Newtown, PA: Law School Admission Council.

Nandakumar, R. y Roussos, L. A. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29*, 177–200.

Pommerich, M., Segall, D.O., y Moreno , K.E. (2009). The nine lives of CAT-ASVAB: Innovations and revelations. In D. J. Weiss (Ed.)*, Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieval [13-04-2011] from www.psych.umn.edu/psylabs/CATCentral/

Revuelta, J., y Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Roussos, L. A., y Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*, 215–230.

Roussos, L. A., y Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*, 215–230.

Shealy, R., y Stout, W. F. (1993b). A model-based standardization approach that Separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Stark, S., Chernyshenko, O. S., y Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1291-1306.

Steinberg, L., y Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402-415.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools.* (Research Report ETS-RR-94-5). Princeton, NJ:Educational Testing Service.

Sympson, J. y Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association,* (pp. 973-977). San Diego, CA:Navy Personnel Research and Development Center.

Thissen, D., Steinberg, L., y Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., y Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer y H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Veldkamp, B. y van der Linden, W., (2007). Designing item pools for adaptive testing. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized Adaptive Testing. Theory and Practice* (págs. 331-352). Boston, MA: Kluwer Academic Publishers. 2a edicion.

Walker, C., Beretvas, N. y Ackerman, T. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. *Applied measurement in education, 14*, 3-16.

Walter, J., Postma, K., McHugh, G., Rush, R., Coyle, B., Strong, V., Sharpe, M. (2007). Performance of the Hospital Anxiety and Depression Scale as a screening tool for major depressive disorder in cancer patients. *Journal of Psychosomatic Research. 63*, 83-91.

Wang, W. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education, 72*, 221-261.

Wang, W., y Yeh, Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Way, W.D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice. 17*, 17-27.

Wells, C., Subkoviak, M., Serlin, R., (2002). The effect of item parameter drift on examinee ability estimates. *Applied psychological measurement. 26*. 77-87.

Woods, C.M. (2008a). Empirical selection of anchors items. *Applied Psychological Measurement, 33*, 42-57.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized Adaptive Testing. Theory and Practice* (págs. 331-352). Boston, MA: Kluwer Academic Publishers.

Zwick, R. (2007). The investigation of Differential Item Functioning in adaptive test. En W. J. van der Linden y C. A. W. Glas (Ed.), *Elements of adaptive testing.* (págs. 331-352). New York: Springer.

Zwick, R., Thayer, D. y Wingersky, M. (1993). *A simulation study of methods for assessing differential item functioning in computer-adaptive tests.* (ETS Research Report 93-11). Princeton, NJ: Educationl Testing Service.

Zwick, R., Thayer, D. y Wingersky, M. (1994a) A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121–140.

Zwick, R., Thayer, D. T. y Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32*, 341–363.

Capítulo 7

Conclusiones y Discusión General

7.1. Consideraciones Generales

Dos son las fases principales para detección del DIF en ítems usados en TAIs. En la fase pretest, ítems cuyos parámetros se desconocen son incorporados al banco. Simultáneamente, se calibran y se analiza su posible DIF. La segunda fase es la de uso operativo de estos ítems. Tal y como hemos argumentado en capítulos previos, el uso del test de razón de verosimilitudes de la TRI (IRT LRT) parece la elección natural de detección de DIF en el contexto de los TAIs, ya que ambas aplicaciones están sustentadas en el mismo modelo psicométrico, la TRI. En el capítulo 3 se describieron los problemas para el análisis de DIF en TAIs, tanto los comunes para las dos fases como los específicos para cada una de ellas.

Hasta el momento, sólo un estudio ha usado el IRT LRT para el análisis de DIF en TAIs, específicamente para ítems pretest (Lei, Chen y Yu, 2006). En este estudio, el modo de aplicar el IRT LRT viene marcado por las características y limitaciones del *software* empleado, IRTLRDIF (Thissen, 2001). La estimación en este programa es calibración con parámetros libres, por lo que la alta tasa de respuestas faltantes en la matriz de datos se convierte en un problema por la inestabilidad de los resultados y su frecuente no convergencia. La necesidad de recalibrar la totalidad de los ítems operativos alarga el tiempo de cómputo. Y, por último, el *software* impide incluir todos los ítems operativos como test de anclaje. Lei et al. (2006) proponen: (a) imputar las respuestas faltantes tomando el nivel de rasgo estimado como nivel de rasgo real; y (b)

reducir la longitud del test de anclaje. En el estudio 1 se ha demostrado que no resulta una estrategia óptima.

Sin embargo, tal y como mostramos en el capítulo cuarto, el problema de la calibración online (y la aplicación del IRT LRT es un caso de calibración online) no es nuevo. Uno de los modos más efectivos para resolver la calibración online es mediante métodos de calibración con parámetros fijos (CPF; Ban et al., 2001; Ban et al., 2002; Kim, 2006). Con CPF desaparece la necesidad de recalibrar los ítems operativos, para los que existe poca información. Además se añade la ventaja de que los parámetros de los ítems pretest y los niveles de rasgo se obtienen en la métrica de los ítems originales. Nuestra propuesta ha sido adaptar estos procedimientos al contexto multigrupo, lo que facilita la aplicación del IRT LRT. Para ello, se ha empleado como software de calibración el ICL (Hanson, 2002), de mayor flexibildad que IRTLRDIF.

Los resultados obtenidos en la aplicación del IRT LRT adaptado para la detección del DIF en ítems pretest (Estudio 1) e ítems operativos (Estudio 2) son prometedores. En ambos trabajos se estudian las tasas de error Tipo I, la potencia, el tamaño del efecto y el del impacto. En el Estudio 1 se contrastaron varios métodos CPF (OWU-OEM, OWU-MEM y MWU-MEM). La principal conclusión, en relación a los métodos CPF, es que el mejor método en el contexto de la calibración online para el caso de un único grupo, MWU-MEM, lo es también en este contexto. Esto se debe a que con MWU-MEM la información del ítem pretest se utiliza para actualizar tanto la distribución posterior como la previa. En cualquier caso, las diferencias entre los métodos CPF fueron pequeñas.

Por el contrario, la imputación de respuestas propuesta por Lei et al. (2006) introduce importantes sesgos en la estimación de los parámetros de los ítems, del tamaño del efecto y del impacto, especialmente cuando existe impacto y el TAI es corto. Estos sesgos se deben a que la estimación del nivel de rasgo con el que se realiza la imputación, contiene error de medida. Paradójicamente, la sobreestimación del tamaño del efecto en presencia de DIF llevó a una mayor potencia de la técnica de imputación lo que alerta sobre la necesidad de incluir indicadores de recuperación de tamaño del DIF en los estudios de simulación. Esto es importante, pues la decisión de descartar un ítem se basa generalmente en criterios que combinan la significación estadística con la medida de tamaño del efecto.

En el Estudio 2 se compara el IRT LRT (con CPF) y el CATSIB para la detección del DIF unidireccional en ítems operativos. Este estudio complementa al

primero, ya que nos permite contrastar: (a) los posibles efectos negativos que pudiera tener la restricción de rango en el rasgo de los que responden al ítem analizado; y (b) la robustez de las técnicas ante la presencia de contaminación en el banco. Además, permitió estudiar el funcionamiento de CATSIB para la detección del DIF en ítems operativos, lo que no había sido investigado previamente.

De nuevo, en presencia de contaminación, IRT LRT proporciona potencias adecuadas en presencia de DIF grande, especialmente en la condición de muestras de tamaño razonable (500 evaluados por grupo). Por otro lado, se mantuvo el control de la tasa de error Tipo I en todas las condiciones. Por tanto, puede considerarse que esta técnica es robusta a la presencia de contaminación en las condiciones simuladas. Por el momento, es difícil evaluar si esto se debe a la presentación adaptativa de los ítems, al uso de las técnicas CPF de calibración o a una combinación de ambas circunstancias. Otra posibilidad es que la proporción de ítems con DIF sea insuficiente. En tests fijos, IRT LRT ha mostrado ser bastante robusto cuando no más del 12% de ítems del subtest de anclaje tiene DIF (Cohen, Kim y Wollack, 1996).

La aplicación de CATSIB requirió decisiones en relación a si aplicar la corrección por regresión (ya que la variable de igualación, θ_{TAI}, incluye al ítem) o al número de intervalos de agrupación para el cálculo de β. En ausencia de contaminación, CATSIB proporcionan un control adecuado de la tasa de error Tipo I, especialmente cuando se aplica la corrección (cuando no se aplica se produce una ligera sobreestimación en presencia de impacto). La presencia de muestras de tamaño distinto (750-250) tampoco produjo una mayor tasa de error Tipo I, lo que difiere de lo encontrado en estudios previos con ítems pretest. Nandakumar y Roussos (2001, 2004) y Lei et al. (2006), encontraron necesario disminuir el número de intervalos a 10 cuando las muestras eran distintas (900-100), para reducir la tasa de error Tipo I. La diferencia de resultados se debe probablemente a la mayor restricción de rango en el rasgo y el mayor solapamiento de las distribuciones de las personas que responden a ítems operativos. Otra posibilidad puede ser que el tamaño del grupo focal en esos estudios (100) resulte más extremo. En cualquier caso, puesto que en esta condición no obtuvimos resultados distintos entre utilizar igual o diferentes tamaños muestrales, se consideró adecuada la elección del número mínimo de intervalos (20).

La condición de banco contaminado ofreció resultados menos favorables a CATSIB. En presencia de contaminación se infló la tasa de error Tipo I, especialmente

en presencia de impacto, y sobre todo cuando no se aplicaba la corrección por regresión. Esto dificulta la interpretación de la potencia obtenida con esta técnica.

7.2. Limitaciones de los estudios y futuras investigaciones

En los estudios presentados se han planteado condiciones realistas en la generación de datos en relación a los parámetros de los ítems o las características del algoritmo adaptativo. Sin embargo, convendría tener en cuenta algunos factores que permitirían establecer la generalizabilidad de los resultados:

- En relación con la propuesta de Lei, et al. (2006) de imputar las respuestas y utilizar la calibración con parámetros libres (CPL), se decidió aislar los efectos de la imputación de respuestas del método de estimación y, por lo tanto, se combinó la imputación de respuestas con el mejor método de CPF (MWU-MEM). Esta decisión se tomó por los siguientes motivos: (a) el método de CPL era poco eficiente, requiriendo cuatro veces más tiempo de cómputo que la CPF; y (b) en la CPL se pierde la métrica, por lo que se ponía en riesgo la comparación sobre la recuperación de parámetros y el tamaño del efecto. Futuros trabajos deberían valorar la aplicación de la propuesta de imputación al método CPL, aunque nuestra previsión es que los resultados serán peores.
- Parece necesario utilizar un número mayor de condiciones en algunas variables (p.ej. tamaños muestrales) que nos permitan tener una visión más completa en relación a los trabajos previos.
- Igualmente, las características del banco elegido, si bien realistas, no lo convierten en perfectamente representativo de todos los bancos de ítems empleados en TAIs. Otras distribuciones de parámetros (univariadas y multivariadas), modelos de TRI en los que están basados o tamaños tendrían que ponerse a prueba.
- Tanto el Estudio 1 como el Estudio 2 comparten una serie de supuestos y características que permiten sugerir futuras líneas de investigación para comprobar la eficacia de las diferentes técnicas de detección de DIF ante el incumplimiento de tales supuestos. En este sentido, consideramos que tres características son muy relevantes:

(a) *Distribución normal de las poblaciones de evaluados*. En relación con este supuesto, en TRI se sabe que los parámetros estimados pueden estar sesgados cuando la distribución $f(\theta)$ no es normal y se ajusta un modelo en el que se asume normalidad (De Ayala, 1995; van den Oord, 2005). En el caso de ítems pretest, la distribución de aquellos que responden al ítem a calibrar equivale a la distribución general de examinados. En este caso, el supuesto de normalidad podría ser poco realista cuando los grupos que se comparan tienen niveles muy altos de habilidad (en contextos de selección) o cuando se evalúan constructos relacionados con la patología (Gibbons, Weiss, Kupfer, Frank, Fagiolini, Grochocinski, Bhaumik, Stover, Bock, y Immekus, 2008; Hart, Deutscher, Crane y Wang, 2009).

En el caso de ítems operativos, la distribución de población de evaludos no coincide con la distribución de los examinados que reciben cada ítem. Tanto es así, que la distribución cambia de ítem en ítem. En el Estudio 2, hemos se consideró que la media y desviación típica de los examinados a los que se les presenta el ítem eran estadísticos suficientes para caracterizar la distribución, puesto que hemos asumido que el nivel de habilidad se distribuye normalmente. Muy probablemente no sea así.

La desviación de la normalidad puede tener importantes consecuencias. Recientemente, Woods (2008b) demostró que IRT LRT presentaba tasas de error Tipo I infladas, imprecisiones en los parámetros de los ítems y una sobrestimación de la media y la varianza de la distribución cuando la distribución de ambos grupos era asimétrica y el procedimiento de estimación asumía normalidad. En estos casos, Mantel-Haenszel o CATSIB podrían ser alternativas preferibles, puesto que no tienen supuestos sobre la distribución. Convendría adaptar y comprobar el funcionamiento de las técnicas propuestas por Woods (2008c) para la estimación de la distribución del rasgo sin asumir normalidad.

(b) *Ausencia de contaminación en el test de anclaje.* Al aplicar IRT LRT, se asume que los ítems de anclaje no tiene DIF. En el Estudio 2 se encuentra que la técnica es robusta en las condiciones simuladas, pero se debería explorar, de manera más sistemática, los límites de la técnica. En el contexto de los tests fijos, se sabe que la tasa de error Tipo I se incrementa

en todos los métodos a medida que aumenta el nivel de contaminación en el subtest de anclaje o subtest de igualación. En el Estudio 2 se simuló un 5% de ítems del banco con DIF. Sin embargo, los test adaptados a otras culturas y/o los que se aplican a gran escala, como los TAIs, muestran porcentajes de ítems con DIF altos (Gierl, Gotzmann y Boughton, 2004). Por tanto, deberían valorarse un porcentaje más alto de ítems con DIF en el banco contaminado.

Por otro lado, la naturaleza adaptativa de los TAIs dificulta la valoración exacta de la presencia de contaminación. Mientras que sólo el 5% del banco presenta DIF, esto no equivale a que sólo el 5% de los ítems del test tengan DIF. Nuestro procedimiento de simulación supone concentrar la contaminación entre aquellos ítems con mayor tasa de exposición para el grupo de referencia. Por tanto, (a) la proporción promedio de ítems con DIF en cada test es mayor al 5%; (b) este porcentaje cambia de examinado a examinado; y (c) posiblemente, cambie también la proporción promedio entre condiciones con impacto y sin impacto.

(c) *Regla de selección de ítems empleada*. En ambos estudios, la regla de selección de ítems aplicada es el método progresivo (Revuelta y Ponsoda, 1998). Esta regla está, muy probablemente, entre las mejores reglas disponibles actualmente cuando se consideran simultáneamente criterios de precisión de medida y de seguridad del banco de ítems (Barrada, Olea, Ponsoda y Abad, 2010), pero hay muchas otras reglas disponibles y actuamente implementadas en programas de evaluación mediante TAIs. La alta relevancia del azar en la selección de ítems al comienzo del test que supone el método progresivo tiene algunos efectos en comparación con selección por máxima información, la regla estándar. Primero, una mayor varianza en el nivel de rasgo de los examinados que reciben un ítem, atenuando parcialmente los problemas de restricción de rango comentados anteriormente. Segundo, aquellos ítems con mayor tasa de exposición no son administrados en fases iniciales del test, sino en las fases finales (Barrada, Olea, Ponsoda y Abad, 2009). Por tanto, es probable que el efecto de la contaminación del banco en el Estudio 2 dependa de la regla de selección de ítems empleada. Con máxima información, es más probable que el primer ítem administrado ya presente DIF y que, por tanto,

la selección de los siguientes ítems ya sea subóptima por la presencia de un primer ítem desajustado.

- Debería estudiarse si es posible mejorar los resultados obtenidos con CATSIB considerando:

  (a) El número óptimo de intervalos a utilizar y la forma de discretizar la variable de igualación (Roussos, Nandakumar y Banks, 2006), entendiendo que la forma optima puede depender del tipo de ítem analizado, pretest u operativo (Lei et al., 2006; Nandakumar y Roussos, 2004; Roussos, et al., 2006; Zwick, 1997, Zwick et al., 1993, 1994a, 1994b).

  (b) El uso de la estimación del nivel de habilidad, sin considerar el ítem analizado e incluyendo la corrección por regresión. Este procedimiento resulta más complejo, pero conceptualmente resulta más apropiado y puede proporcionar mejores resultados.

  (c) La posible existencia de circunstancias más favorables al uso de CATSIB. Por ejemplo, la contaminación del test se produce en el Estudio 2 incluyendo ítems que favorecen al mismo grupo, lo que puede ser poco realista (Wang y Yeh, 2003). Aunque esto resulta justificado en una situación de filtración del contenido de los ítems, sería interesante estudiar los efectos de la contaminación en ausencia de funcionamiento diferencial del test por cancelación del funcionamiento diferencial a través de los ítems. Se sabe que la contaminación tiene menos efectos en esas circunstancias en tests fijos pero ese resultado puede no ser generalizable a los TAIs.

  (d) La posibilidad de usar criterios combinados (significación estadística y tamaño del efecto) para decidir sobre la presencia de DIF en un ítem. Los criterios combinados pueden ser más robustos, considerando que las estimaciones del tamaño del efecto sean apropiadas.

- Es muy común, en las simulaciones de TAIs, considerar que los parámetros de los ítems son conocidos, lo que constituye una práctica poco realista (Zwick, 2007). En futuros estudios es importante evaluar el efecto de utilizar los

parámetros estimados para la selección de los ítems, la estimación de la habilidad de los evaluados o para fijar los parámetros en el procedimiento CPF.

Por otro lado, creemos que existen varios campos prometedores en el estudio del DIF en TAIs:

- En relación al efecto de la contaminación del test, existe gran cantidad de investigación referente a los efectos de la contaminación y las estrategias de depuración del test de anclaje se consideran una parte importante en el estudio del DIF (French y Finch, 2008; González-Betanzos y Abad, en prensa; Lopez-Rivas, Stark y Chernyshenko, 2009; Stark et al., 2006; Wang y Yeh, 2003; Woods, 2008a). Convendría estudiar los efectos de la depuración del test y evaluar las estrategias menos costosas de depuración.
- Debe profundizarse en el estudio de las medidas de tamaño del efecto del DIF obtenidas en el contexto de los TAIs. En primer lugar, el tamaño del efecto del DIF obtenido en la aplicación adaptativa puede diferir del obtenido en la aplicación pretest y los investigadores deberían estar prevenidos sobre ello. Por otro lado, estudios previos (Zwick, 2007) sugieren que si el ítem analizado se estudia adaptativamente, puede haber otras medidas preferibles a β (las que sean análogas a $\mathrm{Ln}(\alpha_{MH})$). Convendría estudiar esto en el contexto de la TRI.
- Urge proponer indicadores que permitan extraer, sin necesidad de realizar un proceso intensivo de simulación, la contribución de cada ítem al funcionamiento diferencial del TAI.
- Se hace necesario el desarrollo de software amigable para la aplicación del IRT LRT en el contexto de los TAIs. El uso de esta técnica en tests fijos estuvo obstaculizado durante mucho tiempo por la complejidad de los análisis necesarios para aplicarla. No fue hasta la aparición del IRTLRTDIF v2.0 (Thissen, 2001), que realiza las comparaciones de modelos anidados de forma automática, que aumentó la frecuencia de uso de esta técnica para la detección del DIF en tests fijos. Ya hemos mencionado que la versión actual tiene serias limitaciones, lo que dificulta el uso en el ámbito de los TAIs. Por otro lado, la implementación de la prueba IRT LRT en programas tales como MULTILOG (Thissen, 1991), BILOG-MG (Zimowski, Muraki, Mislevy y Bock, 2003) o ICL

(Hanson, 2002) puede resultar compleja o imposible. Por ejemplo, aunque MULTILOG (Thissen, 1991) tiene menos limitaciones respecto al número de ítems analizables, se fija automáticamente la varianza de los grupos, lo que resulta inapropiado para los procedimientos CPF. La aplicación de BILOG-MG 7.0, requiere una cuidadosa elección de las opciones y da resultados inadecuados en ausencia de las decisiones correctas que pueden ser difíciles de tomar (Kim, 2006). La opción de utilizar ICL (Hanson, 2002) tiene la ventaja de que es el programa en el que se han implementado tradicionalmente los procedimientos de CPF (Kim, 2006), tiene una gran versatilidad de opciones para la estimación de los parámetros de los ítems (p.ej., en la elección de las distribuciones previas) y se trata de software abierto. Sin embargo, para facilitar su uso resultaría recomendable programar un interface que automatice la realización de todas las comparaciones entre modelos anidados, que facilite la aplicación de estas técnicas por usuarios no expertos.

7.3. Conclusiones

Existen un conjunto de criterios estándar para el análisis del DIF en los tests convencionales. Sin embargo, dichos criterios deben adecuarse a las características que presenta un TAI. En esta tesis hemos mostrado que el procedimiento IRT LRT resulta fácilmente aplicable cuando se extienden los métodos CPF al contexto de los modelos multigrupo. La adaptación de IRT LRT proporciona resultados adecuados en relación a la tasa de error Tipo I, la potencia, la estimación del tamaño del efecto y la estimación del impacto

Referencias

Abad, F. J., Olea, J., Aguado, D., Ponsoda, V., y Barrada, J. R. (2010). Deterioro de parámetros de los items en tests adaptativos informatizados: *Estudio con eCAT. Psicothema, 22*, 340-347.

Abad, F.J, Olea, J., Real, E. y Ponsoda, V. (2002). Estimación de habilidad y precisión en tests adaptativos informatizados y tests óptimos. Un caso práctico. *Revista Electrónica de Metodología Aplicada, 7*, 1, 1-20.

AERA, APA, NCME. (1999). *Standards for educational and psychological testing*. Washington, D.C: American Psychological Association.

Ankenmann, R., Witt, E. y Dunbar, S. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*, 277-300.

Backhoff, E., Ibarra, M.,Rosas, M. y Larrazolo, N. (1999). Sistema de evaluación informatizada para el ingreso a la universidad. En Olea, V. Ponsoda, y G. Prieto (Ed.), *Test informatizados: Fundamentos y aplicaciones*. (pp. 325-342). Madrid: Pirámide.

Baker, F.B., y Kim, S.H. (2004). *Item response theory: Parameter estimation techniques (2nd ed.)*. New York: Marcel Dekker.

Ban, J.C., Hanson, B.A., Wang, T., Yi, Q., y Harris, D.J. (2001). A comparative study of on-line pretest-item calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191-212.

Ban, J.C., Hanson,B.A., Yi, Q., Harris, D.J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement, 39*, 207-218.

Barrada, J. R. (en prensa). Test adaptativos informatizados: una perspective general. *Anales de Psicología.*

Barrada, J. R., Abad, F. J., y Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21*, 318-325.

Barrada, J. R., Olea, J., Ponsoda, V., y Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

Barrada, J. R., Olea, J., Ponsoda, V., y Abad, F. J. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology, 5*, 7-17.

Barrada, J. R., Olea, J., Ponsoda, V. y Abad, F. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Psychological Measurement, 34*, 438-452.

Barrada, J. R., Veldkamp, B. P., y Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33*, 58-73.

Bock, R.D., y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bolt, D. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement. 37*, 307-327.

Bolt, D. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 2*, 113–141.

Bolt, D., y Gierl, M. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement, 43*, 313-334.

Bolt, D. y Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting sibtest detection procedure. *Behaviormetrika*, J*ournal of the Behaviormetric Society of Japan, 23*, 67-95.

Borsboom, D., Romeijn, J. y Wicherts, J. (2008). Measurement invariance versus Selection Invariance: Is Fair Selection Possible? *Psychological Methods. 13*, 75–98.

Camilli, G., y Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage publications

Candell, G. y Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253–260.

Chang, H. H.. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In David Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (págs. 117-133). Thousand Oaks, CA: Sage Publications.

Chang, H. H., y Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H. H., y Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, H. H., y Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387-398.

Chang, H., y Ying, Z. (2006, August). *Making item selection more efficient in computerized adaptive testing*. Paper presented at the Joint Statistical Meeting, Seattle, WA. Paper presented at the Joint Statistical Meeting, Seattle, WA

Chen, S. Y., Ankenmann, R. D., y Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.

Cheng, Y., y Chang. H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369-383.

Clauser, B., Mazor, K.., y Hambleton, R. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269–279.

Davey, T., y Nering, N. (2002). Controlling item exposure and maintaining item security. En C. N. Mills, M. T. Potenza, J. J. Fremer, y W. C. Ward, (Ed.), *Computer-based testing: Building the foundation for future assesments* (págs. 165-191). Mahwah, NJ: Lawrence Erlbaum.

Davey, T., y Parshall, C. G. (1995, Abril). *New algorithms for item selection and exposure control with adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

De Ayala, R. J. (1995, April). *Item parameter recovery for the nominal response model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

DeMars, C. (2009). Modification of the mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of educational and behavioral statistics, 34*, 149-170.

DeMars, C. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and psychological measurement. 70*, 961-972.

Diao, Q., y van der Linden, W. J. (en prensa). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*.

Dodd, B. G. (1990) The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.

Donoghue, J., Holland, P., y Thayer, D. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland y H. Wainer (Ed.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*, 217-233.

Dorans, N., y Holland, P. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland y H. Wainer (Ed.), *Differential item functioning* (pp.35–66). Hillsdale, NJ: Erlbaum.

Dorans, N. y Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in december 1977*: An application of the standardization approach. (Research Report R116). Princeton, NJ: ETS.

Embretson, S. (2004). The second century of ability testing some predictions and speculations. *Measurement, 2*, 1-32

Fidalgo, A. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord.), *Psicometría* (pp. 370-455). Madrid: Universitas.

Fidalgo, A., Mellembergh, G., y Muñiz, J.(2000). Effects of Amount of DIF, Test Length, and Purification Type on Robustness and Power of Mantel-Haenszel Procedures. *Methods of Psychological Research Online. 5*, 43-58.

Fidalgo, A., Mellembergh, G., y Muñiz, J. (1998). Comparación del procedimiento MH frente a los modelos loglineales en la detección del funcionamiento diferencial del los ítems. *Psycothema, 10(1)*, 209-218.

Fidalgo, A., Mellembergh, G., y Muñiz, J. (1999). Aplicación en una etapa, dos etapas e iterativamente de los estadísticos Mantel-Haenszel. *Psicológica, 20*, 227-242.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295

Flaugher, R. (2000). Item pools. En Wainer, H. (Ed.), *Computerized adaptive testing: a primer*. 2º ed. (pp.37-59).Hillsdale, NJ:LEA. pp. 335

French, B., y Finch, H. (2008). Multigroup confirmatory factor analysis: locating the invariant referent sets. *Structural Equation Modeling, 15*, 96-113.

French, B., y Maller, S. (2007). Iterative Purification and Effect Size Use With Logistic Regression for Differential Item Functioning Detection. *Educational and Psychological Measurement, 67*, 373-393.

Gibbons, R., Weiss, D., Kupfer, D., Frank, E., Fagiolini, A., Grochocinski, V., Bhaumik, D., Stover, A., Bock, D., y Immekus, J. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361-368.

Gierl, M., Gotzmann, A., y Boughton, K. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17*, 241-264.

Gomez-Benito, J., Hidalgo, M., y Guilera, G. (2010). El sesgo en los instrumentos de medición: test justos. *Papeles del psicólogo: revista del Colegio Oficial de Psicólogos, 31*, 75-84.

Gonzalez-Betanzos, F. y Abad, F., (en prensa). The effects of purification and the evaluation of differential item functioning with the likelihood ratio test. *Methodology.*

Hanson, B. A (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*, 244-253.

Hanson, B. A. (2002). *IRT Command Language* (Version 0.020301). Monterey, CA: Author. (Available at http://sourceforge.net/projects/ssm)

Hart, D., Deutscher, D., Crane, P y Wang, Y. (2009). Differential item functioning was negligible in adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. *Quality Life Research. 18*, 1067-1083.

Harwell, M. y Baker, F. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement. 15*, 375-389.

Hetter, R. D., y Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. En W. A. Sands, B. K. Waters, y J. R. McBride (Ed.), *Computerized adaptive testing: From inquiry to operation* (págs. 141-144). Washington DC: American Psychological Association.

Hidalgo, M. D., y López-Pina, J. A. (2004). DIF detection and effect size: A comparison between logistic regression and Mantel-Haenszel variation. *Educational and Psychological Measurement, 64*, 903–915.

Hidalgo, M., y Gomez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker y B. McGaw (Ed.), *International Encyclopedia of Education* (3rd edition). USA: Elsevier: Science & Technology.

Holland, P. (1985). On the study of differential item performance without IRT. *Proceedings of the Military Testing Association*. October.

Holland, P.y Thayer, D. (1985). *An alternate definition of the ETS scale of item difficulty*. (Research Report No. RR-85-43). Princeton, NJ: Educational Testing Service.

Holland, P. y Thayer, D. (1988). Differential item performance and the Mantel–Haenszel statistic. En: H. Wainer y H. Braun (Ed.), *Test Validity*, Hillsdale: Erlbaum, pp. 129–145.

Holland, P. W. y Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Hornke, L.F. (2000). Item response times in computerized adaptive testing. *Psicológica, 21(1-2)*, 175-189.

Hsu, Y., Thompson, T.D., y Chen, W.H. (1998). *CAT item calibration*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Hulin, C.L., Drasgow, F. y Parsons, C.K. (1983). *Item response theory: Application to psychological measurement*. Homewood, Illinois: Dow Jones/Irwin.

International Test Commission (ITC). (2005). *Guidelines on computer-based and internet-delivered testing*. Accedido el 19 de mayo, 2011, de http://interestcom.org/

Jiang, H., y Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*, 291-322.

Jodoin, M.G., y Gierl, M.J. (2001). Evaluating Type 1 error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43(4)*, 355-381.

Kim, S.H., y Cohen, A.S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*, 131-143.

Kingsbury, G. G., y Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

Klockars, A., y Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement, 45*, 271-285.

Lautenschlager, G. J., Flaherty, V. L. y Park, D. G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Educational and Psychological Measurement, 54*, 21-31.

Lei, P.-W., Chen, S.-Y. y Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*, 245–264.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. En P. W. Holland y H. Wainer (Ed.), *Differential item functioning* (pp. 317-319). Hillsdale, NJ:Lawrence Erlbaum Associates, Inc.

Li, H. y Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.

Li, Y. H., y Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42*, 245-269.

Linn, R., Levine, M., Hastings, N., y Wardrop, J. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

Lopez-Rivas, G., Stark, S., y Chernyshenko, O. (2008). The effects of referent item parameters on Differential Item Functioning detection using the Free Baseline Likelihood Ratio Test. *Applied Psychological Measurement, 33*, 251-265.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Mazor, K. M., Clauser, B. E., y Hambleton, R. K. (1994). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 54*, 284–291.

McLachlan, G. J., y Krishnan, T. (1997). *The EM algorithm and extensions*. New York: John Wiley & Sons.

Miller, M. D., y Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*, 381-388.

Millsap, R. y Everson, H. (1993). Methodology review: statistical approaches to assessing measurement bias, *Applied Psychological Measurement, 17*, 297–334.

Mislevy, R. J., y Bock, R. D. (1985). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. En D. J. Weiss (Ed.), *Proceedings of the 1982 item response theory and computerized adaptive testing conference* (pp. 189-202). Minneapolis MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Mislevy, R.J. y Wu, P. (1996). *Missing responses and IRT ability estimation: omits, choice, time limits, and adaptive testing* (Research Report RR-96-30-ONR). Princeton, NJ:ETS.

Monahan, P. O., y Ankenman, R. (2010). Alternative matching scores to control type I error of the Mantel Haenszel procedure for DIF in dichotomously scored items conforming to 3PL IRT and Nonparametric 4PBCB models. *Applied Psychological Measurement, 34*, 193-210.

Monahan, P. O., McHorney, C. A., Stump, T. E., y Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*, 92-109.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Sheally-Stout´s test for DIF. *Journal of Educational Measurement, 30*, 293-311.

Nandakumar, R., y Roussos, L. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive test*. Law School Admission Council Computerized Testing Report. LSAC research Report Series.

Nandakumar, R. y Roussos, L. A. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29*, 177–200.

Narayanan, P. y Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Narayanan, P. y Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Olea, J., Abad, F. J., Ponsoda, V., Barrada, J. R., y Aguado, D. (En prensa). eCat-Listening: Diseño y estudio de las propiedades psicométricas de un test adaptativo informatizado para la evaluación del nivel de comprensión auditiva de la lengua inglesa. *Psicothema*.

Olea, J., Abad, F. J., Ponsoda, V., y Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: Diseño y comprobaciones psicométricas. *Psicothema, 16*, 519-525.

Olea, J., Ponsoda, V., Revuelta, J., Hontangas, P. y Suero, M. (1999). Investigación en test adaptativos informatizados. En J. Olea, V. Ponsoda y G. Prieto (Ed). *Test informatizados. Fundamentos y Aplicaciones*. (pp. 163-185). Madrid. Pirámide.

Olea, J. y Ponsoda, V. (2003). *Tests adaptativos informatizados*. Madrid:UNED, Colección Aula Abierta.

Olea, J., Revuelta, J., Ximénez, C. y Abad, F.J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive test. *Psicologica, 21*, 157-173.

Oshima, T. y Miller, D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*, 237-248.

Osterlind, S., Everson, H. (2009). Differential item functioning. *Series Quantitative Applications in the social sciences*. Vol. 161, pp. 87.

Parshall, C.G. (1998). *Item development and pretesting in a computer-based testing environment.* Paper presented at he colloquium Computer-based testing: Building Foundation for Future Assessments, Philadelphia, PA.

Pei, L. y Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood Ratio for DIF detection. *Applied Psychological Measurement, 34*, 453-456.

Penfield, R. y Camilli, G. (2007). Differential Item Functioning and Item Bias. En C.R. Rao y S. Sinharay (Ed.), *Handbook of Statistics: Psychometrics*. vol. 26, pp. 125-168. Elsevier: Amsterdam, The Netherlands.

Ponsoda, V., Revuelta, J., Hontangas, P., y Suero, M. (1999). Investigación en test adaptativos informatizados. En J. Olea, V. Ponsoda, y G. Prieto (Ed.), *Test informatizados: fundamentos y aplicaciones*. (pp.163-185). Madrid: Pirámide.

Powers, D. E. y O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment, 1,* 153–173.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Raju, N. S., van der Linden, W.,J y Fleer, P.F (1992). *An IRT-based internal measure of test bias with applications for differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Raju, N. S., van der Linden, W.J., y Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Rebollo, P., García-Cueto, E., Zardaín, P. C., Cuervo, J., Martínez, I., Alonso, J., Ferrer, M., y Muñiz, J. (2009). Desarrollo del CAT-Health, primer test adaptativo informatizado para la evaluación de la calidad de vida relacionada con la salud en España. *Medicina Clínica, 133*, 241-251.

Revuelta, J., y Ponsoda, V. (1997). Una solución a la estimación inicial en los test adaptativos informatizados. *Revista electrónica de metodología aplicada, 2(2)*,16.

Revuelta, J., y Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Rogers, H. y Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement. 17*, 105-116.

Roussos, L. A., y Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*, 215–230.

Roussos, L. A., Schnipke, D. L., y Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*, 293–322.

Roussos, L., Nandakumar, R., Banks, J. (2006). *Theoretical formula for statistical bias in CATSIB DIF estimator due to discretization of the ability scale.* (Research report). Newtown, PA: Law School Admission Council.

Rubio, V., y Santacreu, J. (2003). *TRASI. Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción como factores de la habilidad intelectual general*. Madrid: TEA ediciones.

Rudner, L., Geston, P., Knight, D. (1980). Biased item detection techniques. *Journal of Educational Measurement Statistics, 5*, 213-233.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*, 439-460.

Shealy, R., y Stout, W. F. (1993a). An item response theory model for test bias. In P. W. Holland y H. Wainer (Ed.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Erlbaum.

Shealy, R., y Stout, W. F. (1993b). A model-based standardization approach that Separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shepard, L., Camilli, G. y Williams, D. (1984). Accounting for statistical artifacts in item bieas research. *Journal of Educational Statistics, 9*, 93-128.

Shepard, L., Camilli, G. y Williams, D. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.

Stark, S., Chernyshenko, O. S., y Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1291-1306.

Steinberg, L., y Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402-415.

Steinberg, L., Thissen, D. y Wainer, H. (1990). Validity. En Wainer (Ed.), *Computer adaptive testing: A primer*. (Capítulo 8, pp. 185-230). Hillsdale, NJ: Lawrence Erlbaum.

Stocking, M.L., (1988). *Scale drift in on-line calibration* (Research Report 88-28). Princeton, NJ: ETS.

Stocking, M. L, y Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Stocking, M. L., y Lewis, C. L. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Stocking, M. L., y Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. En W. J. van der Linden y C. A. W. Glas (Ed.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic.

Swaminathan. H., y Rogers, H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory [Computer software and manual]*. Chicago, IL: Scientific Software International.

Thissen, D. (2001). *IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software and manual]*. Chapel Hill: L.L. Thurstone Psychometric Laboratory, University of North Carolina.

Thissen, D., Steinberg, L., y Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., y Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer y H. Braun (Ed.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum..

van den Oord, E. J. C. G. (2005). Estimating Johnson curve population distributions in MULTILOG. *Applied Psychological Measurement, 29*, 45-64.

van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23*, 21-29.

van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. En W.J. van der Linden y C. A. W. Glas (Ed.), *Computerized adaptive testing: Theory and practice* (págs. 27-52). Boston, MA: Kluwer Academic Publishers.

van der Linden. W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249-265.

van der Linden, W. J. (2005). A comparison of item-selection method for adaptive tests with content constraints. *Journal of Educational Measurement, 45*, 283-302.

van der Linden, W. J., y Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273-291.

van der Linden, W. J., y Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398-418.

Veerkamp, W. J., y Berger, M. P. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

Wainer, H. (2000a). *Computerized adaptive testing: A primer*. Hillsdale, NJ: LEA. (1ª edición: 1990).

Wainer, H., y Mislevy, R.J. (1990). Item response theory, item calibration, and profiency estimation. En Wainer (Ed.), *Computer adaptive testing: A primer*. (Capítulo 4, pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.

Walker, C., Beretvas, N. y Ackerman, T. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. *Applied measurement in education, 14*, 3-16.

Wang, T. y Vispoel, W. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of educational measurement. 35*, 109-135.

Wang, W., y Yeh, Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Welkenhuysen-Gybels, J. (2004). The performance of some observed and unobserved conditional invariance techniques for the detection of differential item functioning. *Quality and Quantity, 38*, 681-702.

Woodruff, D.J., y Hanson, B.A.(1996). *Estimation of ítem response models using the EM algorithm for finite mixtures* (ACT Research Report 96-6). Iowa City, IA: ACT, Inc.

Woods, C. (2008a). Empirical selection of anchors items. *Applied Psychological Measurement, 33*, 42-57.

Woods, C. (2008b). Likelihood-ratio differential item functioning testing: Effects of nonnormality. *Applied Psychological Measurement, 32*, 511-526

Woods, C. (2008c). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement, 68*, 571-586

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland, y H. Wainer (Ed.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum.

Zimowski, M., Muraki, E., Mislevy, R., y Bock, R. (2003). BILOG-MG (version 7.0). In M. du Toit (Ed.), *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* (Chap. 2, pp. 24–256). Lincolnwood, IL: Scientific Software International.

Zumbo, B. D., y Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). Prince George, Canada: University of Northern British Columbia

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 85–197.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. En W. J. van der Linden y C. A. W. Glas (Ed.), *Computerized Adaptive Testing. Theory and Practice* (págs. 221-243). Boston, MA: Kluwer Academic Publishers.

Zwick, R. (2007). The investigation of Differential Item Functioning in adaptive test. En W. J. van der Linden y C. A. W. Glas (Ed.), *Elements of adaptive testing.* (págs. 331-352). New York: Springer.

Zwick, R. y Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel–Haenszel DIF analysis to computer-adaptive tests. *Applied Psychological Measurement, 26,*57–76.

Zwick, R. y Thayer, D. T. (2003, August). *An empirical Bayes enhancement of Mantel–Haenszel DIF analysis for computer-adaptive tests* (Computerized Testing Report No. 98-15). Newtown.

Zwick, R., Thayer, D. T. y Lewis, C. (1997) *An investigation of the validity of an empirical Bayes approach to Mantel–Haenszel DIF analysis* (ETS Research Report No. 97-21). Princeton,NJ: Educational Testing Service.

Zwick, R., Thayer, D. T. y Lewis, C. (1999). An empirical Bayes approach to Mantel–Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1–28.

Zwick, R., Thayer, D. T. yWingersky, M. (1993) *A simulation study of methods for assessing differential item functioning in computer-adaptive tests.* (ETS Research Report 93-11). Princeton, NJ: Educationl Testing Service.

Zwick, R., Thayer, D. T.y Wingersky, M. (1994a) A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121–140.

Zwick, R., Thayer, D. T. y Wingersky, M. (1994b) *DIF analysis for pretest items in computeradaptive testing* (ETS Research Report 94-33). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T. y Wingersky, M. (1995). Effect of Rasch calibration on ability and DIFestimation in computer-adaptive tests. *Journal of Educational Measurement, 32*, 341–363.